DOCUMENT RESUME

ED 065 225 95 RC 006 287

TITLE Bilingual Testing and Assessment, Proceedings of Bay Area Bilingual Education League (BABEL) Workshop and Preliminary Findings, Multilingual Assessment Program (Berkeley, California, January 27-28, 1969).

INSTITUTION Bay Area Bilingual Education League, Berkeley, Calif.; Multilingual Assessment Program, Stockton, Calif.

SPONS AGENCY Office of Education (DHEW), Washington, D.C.

PUB DATE 28 Jan 72

NOTE 122p.

EDRS PRICE MF-$0.65 HC-$6.58

DESCRIPTORS Biculturalism; *Bilingual Education; *Conference Reports; Culture Free Tests; Intelligence Tests; *Minority Groups; Norm Referenced Tests; Tables (Data); *Testing; *Test Interpretation; Test Reviews

ABSTRACT

The results and proceedings of the first annual Bilingual/Bicultural Testing and Assessment Workshop, held in Berkeley, California, on January 27-28, 1972, are presented in this publication. Approximately 150 bilingual psychologists and evaluators, educators working in bilingual/bicultural programs, and community representatives from California and Texas attended. Evaluations were made and the summaries are included of 8 tests used extensively in bilingual programs: the Wechsler Intelligence Scale for Children, the Comprehensive Tests of Basic Skills, the Cooperative Primary, the Lorge-Thorndike, the Inter-American Series--General Ability, the Culture-Fair Intelligence Test, the Michigan Oral Production Test, and the Peabody Picture Vocabulary Test. Also included in this publication are (1) an overview of the problem of assessment and evaluation in bilingual education, (2) a professional critique of the Inter-American series by Dr. Barbara Havassy, (3) a brief description of a Criterion Referenced System developed by Eduardo Apodaca, and (4) an article by Dr. Edward A. DeAvila discussing some of the complexities involved in testing and assessment of bilingual/bicultural children. (NQ)

BILINGUAL TESTING AND ASSESSMENT

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

RECEIVED
AUG 14 1972

ED 065225

BABEL

BAY AREA BILINGUAL EDUCATION LEAGUE

PROCEEDINGS

OF

BABEL WORKSHOP (January 27-28, 1972, Berkeley)

AND

PRELIMINARY FINDINGS

MULTILINGUAL ASSESSMENT PROGRAM

RC006287

RECEIVED
AUG 14 1972
NMSU
E.R.I.C.

Preface

This publication is the fruition of a joint effort by the Bay Area Bilingual Education League, Berkeley, and the Multilingual Assessment Program, Stockton. I wish to acknowledge Dr. Rene Cardenas, Director of BABEL and Mr. Joe R. Ulibarri, Director of the Multilingual Assessment Program, for their fine cooperation, interest, and support of this effort.

It was the intent of this publication to share the results and proceedings of the first annual Bilingual/Bicultural Testing and Assessment workshop. Enclosed herewith the reader will find a test by test summary of the discussion and conclusions reached during the workshop sessions. The reader will find that this represents a lay effort to find a solution to one of the most pressing problems in Bilingual Education, that of Assessment and Evaluation. An overview of this problem is also presented so that the reader might gain additional insight into the processes that led up to the present "state of the art" in Bilingual/Bicultural Educational Evaluation. We are grateful to be able to include a professional critique of the Inter-American Series by Dr. Barbara Havassy, consultant to the Multilingual Assessment Program, Stockton, even though it was not an official part of the assessment workshop. Her work, entitled "A Critical Review of the New Inter-American Series" should make a timely

and valuable contribution to those projects contemplating use of this instrument.

A brief description of a Criterion Referenced System developed by Eduardo Apodaca is also presented and stimulates the appetite for further investigation and experimentation with this methodology.

Dr. Edward A. DeAvila's original presentation, "Testing Tonterias" has been graciously expanded to present additional considerations for our readers. "Some Cautionary Notes on Attempting to Adapt I.Q. Tests for use with Minority Children and a Neopiagetian Approach to Intellectual Assessment: Partial Report of Preliminary Findings" spells out more clearly some of the complexities involved in testing and assessment of Bilingual/Bicultural children.

This is hopefully only a beginning effort by Title VII projects to deal with an area where there are far too few experts and far too many novices attempting to tackle problems that are "Anglo" created, perpetuated and rewarded. As we continue the struggle it is imperative that the fruition of our efforts be shared. In this regard I welcome your comments and suggestions and promise to continue this effort in the San Francisco Bay Area.

Requests for further information on this subject may be addressed to Mr. Joe Ulibarri, Director, Multilingual Assessment Program, 1111 No. El Dorado, Stockton, California, or to me at 1414 Walnut Street, Berkeley, California 94709.

Olivia Martinez
Bay Area Bilingual Education League

June, 1972

BABEL
BAY AREA BILINGUAL EDUCATION LEAGUE
1414 WALNUT STREET
BERKELEY, CALIF. 94709

CONTRIBUTORS

Olivia Garcia Martinez received her M.S.W. in School Social Work from the University of California, Berkeley. She received an advanced credential and training in School Psychology from the California State College, Hayward, and is presently the Coordinator of Testing and Evaluation for the Bay Area Bilingual Education League.

Dr. Barbara Havassy received her Ph.D in Social Psychology from the University of Colorado. She is presently a consultant to the Multilingual Assessment Program in Stockton.

Eduardo A. Apodaca is the Director of Project Hacer Vida, Title VII Bilingual Education program, Office of Riverside County Superintendent of Schools, Coachella Valley Branch.

Dr. Edward A. DeAvila received his Ph.D from York University. He is a Developmental Psychologist and the current Research Director for the Multilingual Assessment Program in Stockton.

Foreword

by Olivia G. Martinez

Bilingual Education originated in 1967 as an amendment to the Elementary and Secondary Education Act. It provided for the development and inclusion of a Bilingual Education program in districts that contained a sizeable number of Spanish-Speaking pupils. Prior to this time school districts that served predominately Spanish-Speaking pupils concentrated on crash "ESL" or "English as a second language" programs that were designed to teach English as soon as possible so that the native Spanish speaker would "function" in a regular classroom.

It is not insignificant that it came after generations and generations of Mexican-American pupils had already been "pushed out" of the educational system. California and the Southwest have long held the distinction of having the largest number of bilingual inhabitants. Bilingual needs have been around for a long time and were documented as early as 1934 when Chicano Educators first made their plea to the Psychological Associations for testing and assessment in one's native tongue.[1] Indeed California's first constitution was written in the Spanish Language! One can hardly turn a corner in California without a glaring reminder of the rich Spanish and Mexican heritage documented throughout the state.

It is sad to note that bilingual education was recognized as a valuable and necessary program for the Southwest only after Congress saw fit to enact legislation to assist the "political refugees" from Castro's Cuba. They drew the very logical conclusion that if they were to welcome and provide for the large influx of Cubans some provision had to be made to accomodate their bilingual needs in education. Thus the first monies were allocated to teach these unfortunate victims of a communist regime the language they would need to know for survival in their new country. Clearly the emphasis was on the acquisition of English. From there it was a fairly simple matter to make the generalization to the southwestern communities who were also seen to be unfit and unprepared because of their language differences to benefit from and contribute to American society.

Today there exists a hodgepodge of programs under the banner of Bilingual Education, but not that many that actually practice what they preach. We at BABEL recognize a program as such only when instruction is offered in the dominant language of the child. The child should be allowed to achieve mastery in his own tongue before introducing a formalized reading program in English. Even then the child should be encouraged actively to continue concept and vocabulary improvement in his first language. Research conducted on Bilingual Education in Canada revealed that pupils who were totally fluent in their first tongue and could read and write their own language had a much easier

time of acquiring a second language fluently and even went on to excell when compared to monolingual peers.[2]

In this country, where pluralistic education has been a vague concept at best, Bilingualism and Biculturalism has been viewed as a handicap! Despite the fact that certain segments of society, as many European societies, have long recognized the desirability of learning two languages, two cultures, etc. Indeed one qualification for entrance to colleges and universities was a foreign language program. Yet, Chicanos have been admonished and discouraged from perpetuating our "ready-made" bilingualism/biculturalism. Nowhere is this "handicap" so evident as in the area of evaluation, testing and assessment. More Chicano children have been labeled, placed, tracked, grouped and guided on the basis of various test scores than on any other single factor in the classroom. While there is no hard data to substantiate this claim, there are considerable statistics to document the failure of the public school system in educating Bilingual, particularly, Chicano children.[3] Sometimes referred to as the "push-out" rate, this well and perhaps overdocumented phenomena in many cases begins with a standardized test of some sort.

Aside from the "routine" testing for special educational needs and placements, an additional phenomena of testing for program effectiveness has emerged as a serious concern to Bilingual/Bicultural Educators. Bilingual Education is of

necessity an innovative program based on an innovative approach to educating all children. How then, can a traditional pre/post test evaluation design, using traditional standardized instruments be expected to effectively evaluate an innovative, multi-component program?

The Bay Area Bilingual Education League (BABEL) has five major components: The Instructional Program, Staff Development, Curriculum and Materials, Higher Education and Media. To expect a standarized test or even a series of tests to document the effectiveness of these highly specialized areas is fallacious to say the least. Yet when school boards and administrators attempt to evaluate a program, particularly with regard to refunding or expansion, invariably one hears the test scores being reported.

California has two required statewide programs for testing pupils in the public schools. They are the California School Testing Program and the testing required under the Miller-Unruh Basic Reading Act of 1965. The California School Testing Program began with a law passed at the 1961 Session of the Legislature for the purpose of revealing the status of California students with respect to the academic skills and content they have acquired. Amended in 1963, the act requires testing with intelligence, achievement and physical performance. The tests adopted for use in the 1969-70 school year are the Lorge-Thorndike Intelligence tests in grades 6 and 12, the Comprehensive Tests of Basic Skills in grade 6, the Iowa Tests of Educational Development in grade 12

and the California Physical Performance Test in grades 6 and 12. Intelligence tests are administered during the months of October and November, achievement tests during the months of October, and physical performance tests during April and May.

The Miller-Unruh Basic Reading Act Testing in grades 1, 2 & 3 was required in connection with a program to improve reading instruction in the primary grades. The Cooperative Primary Reading Test is administered the first 10 school days in May. Test results are reported to the State Department of Education, and one of the uses made of the required testing is in establishing the system of priorities for funding under the Miller-Unruh Basic Reading Act. Also, test results are used for evaluation of reading programs on both the district and State levels.[4]

In addition to a concern over how well California pupils are doing compared to the rest of the nation, the state mandated testing program was seen as a means of prodding districts into revamping instructional procedures. This is apparently accomplished by publishing test scores in local papers where district by district comparisons as well as school by school comparisons could be made. Thus we have a situation where districts and schools are first rewarded for low test scores (qualifying for the Miller-Unruh funds) and then possibly penalized when significant growth is or is not reflected in the scores (evaluation for continued funding). There is considerable evidence to document the inadequacy

of standardized tests for some minority and/or culturally different, bilingual children. If one is dissatisfied with this point of view (based on work done by Dr. Palomares, Dr. Steve Moreno, George Sanchez and others) then he need only refer to the various law suits pending on the misuse of standardized tests results, for Spanish-Speaking children.[5] Yet standardized tests continue to be treated as if they do in fact adequately assess such children. The problem is complex and emotionally charged. If one wants only to know how well Bilingual/Bicultural children perform on standardized I.Q. and achievement tests in comparison to middle class children, and if one wants to know how well minority children can do on a dominant culture value oriented (i.e. how well he can take anglo tests), and if one wants evidence of how implicit functional objectives of various educational programs are failing to serve bilingual/bicultural children, than that is a defensible position. Since a relatively small percentage of people understand testing, test development and statistical inferences, it is well to consider the current use of standardized tests "assume a universality in community of experiences...a test is valid only to the extent that the items of the test are as common to each child tested as they were to the children upon whom the norms were based." The problem as I see it relates to the fact that tests are not administered for the positions described above, or even with those notions in mind. Instead, standardized tests are used as a reflection of the innate,

and potential intelligence of children, as a predictor of future accomplishments (remember the self-fulfilling prophecy), as a device to group and label, and finally as proof of the inadequacy and handicaps a bilingual/bicultural child brings to the educational setting.

Dr. Uvaldo Palomares has described the unique motivational style Chicano youngsters bring to the classroom. He also discusses the concept of positioning and cultural divergence in an attempt to document how standardized I.Q. tests are not fair to Chicano Children.[6] We don't need any more evidence. Most persons knowledgeable about tests and their uses readily agree with George Sanchez's position that the worth of test-results lies in their proper interpretation and in the assistance which such interpretation lends to furthering the educational needs of the pupil. An I.Q. ratio, as such has no value. It is only when that measure is used critically in promoting the best educational interests of the child that it has any worthwhile significance to the educator.[7] Yet test publishers willingly demonstrate how to collapse scores to yield a grade equivalent, I.Q. and percentile rank, that require a tremendous stretch of the imagination to be seen as helpful to the teacher.

I could provide pages and pages of anecdotal material, including several personal experiences that would dramatically illustrate the evils of testing minority students, however, I reject the notion that Minority educators must continue to

perform before our advice is heeded. We know the dangers in using standardized tests is their misuse, the test publishers know it, and many key educators know it. If the State of California, by mandating such tests and allowing their continued misuse is the originator and perpetuator of, say, tracking and labeling, what does that say for California's commitment to equal educational opportunity?

Elsewhere in this publication is a description of a testing and assessment workshop recently hosted in Berkeley by BABEL. This meeting of approximately 150 evaluators, psychologists, and educators was originally conceived because of the dissatisfaction and concern of Chicano, Asian and other bilingual educators, with the continued use of standardized achievement tests and traditional I.Q. tests. As evaluators of Bilingual programs we were particularly concerned about the use of such tests for programmatic evaluation. The problem is multi-dimensional: Bilingual programs need thorough evaluations. We must be able to assess where and how effectively we are going. What is happening to children in our programs that would not otherwise happen to them? As discussed earlier in this paper, there is evidence to suggest that routine testing and assessment of Bilingual/Bicultural children is unhelpful, if not harmful. The simple translation of existing tests is unsatisfactory and merely results in presenting the same unacceptable, culturally biased content in Spanish, (sometimes changing the degree of difficulty

in the process). Development of new bilingual/bicultural instruments is costly, time-consuming and would most likely perpetuate the worn out concept of testing the child and not the system. Besides, there is no one test in existence today that adequately assesses anglo children, let alone the many and various programatic components. Excluding bilingual/bicultural children from existing state and district testing programs suggests a continuation of the "labeling by separation" tendencies we are attempting to destroy.

A recent survey by the Multi-lingual Assessment Program in Stockton revealed the thirteen most commonly used tests in Bilingual Education Projects in California to be as follows:

Culture Fair Intelligence Test
Van Alystyne Picture Vocabulary Test
Peabody Picture Vocabulary Test
Metropolitan Readiness Test
Inter-American Series
Goodenough Draw a Person
Lorge Thorndike*
Stanford Achievement*
Michigan Oral Language Test
Test of Basic Experiences
Metropolitan Achievement Test
Comprehensive Test of Basic Skills*
Cooperative Primary*

*State Mandated Testing Program

Many people have repeatedly criticized these instruments and how unhelpful they are. Few people have actually documented where these tests penalize or harm bilingual/bicultural children, and this was the ambitious task of this first workshop on Testing and Assessment. A second objective was to look at the Criterion Referenced system as an alternative to traditional assessment. The proceedings of this workshop along with the resolutions passed describe how enormously complicated this task was and more than likely attests to the general naivete of persons using such tests. That is, it was only when groups attempted to document the so-called inadequacies of tests that they became truly aware of the intended uses of such instruments and how little they actually knew about them. In several instances, what the author of the test intended, and what the publishers suggested and what the school personnel actually used the tests for were all very different! Few persons took the radical position of categorically condemning all tests for all purposes under any circumstances. However, few could deny that the gross misuses of tests historically and up to the present did warrant such considerations and that perhaps some sort of moratorium might be necessary as an interim measure.

While we were unable to critique all the tests as hoped, in general I felt many people left this workshop more informed and more comfortable in their conviction

that standardized tests should be removed from their position of sanctity and relegated to a more menial place in education, but uncomfortably aware of the fact that the blame for the devastating results labeling has had on bilingual populations does not lie with the test alone; nor will the simple act of discontinuing their use provide the solutions to our dilemma.

In the meantime, then, can we please turn our attention, energy and resources to alternatives to standardized testing, i.e. non-obtrusive measures, behavioral and affective areas and Criterion-Referenced Tests?

THE CRITERION-REFERENCE MODEL

Of the several alternatives presently available to us, the Criterion-Referenced Model appears to be the most promising.

In an article by Rex Jackson entitled "Developing Criterion-Referenced Tests", a definition of Criterion-Referencing is offered as follows:

According to Wang (1969) a "criterion-referenced test is an achievement test developed to assess the presence or absence of a specific Criterion behavior described in an instructional objective". The term appears to have been introduced by Glaser (1963) in a paper in which he distinguishes "criterion-referenced" from "norm-referenced" testing. In the latter, an individual's test performance is interpreted with respect to the performance of other individuals who belong to some specified population. In contrast, the

interpretation of an individuals' performance on a criterion-referenced test is a behavioral statement (or set of statements) that is made without reference to the performance of other individuals.[8] This system has also been referred to as compentency-based or even precision teaching. I feel that essentially they are all the same thing - that is, they all attempt to test what one has been teaching, not what some test developer <u>assumes</u> has been taught.

Two bilingual education programs, one in Indio and the other in Santa Ana, California are currently using such a model and initial indications are very promising.[9]

No one is willing to categorically state that Criterion-Referenced Tests will provide the solutions to all our problems. However, it certainly appears to suit the needs of Bilingual/Bicultural Education more readily than norm-referenced or standardized tests. Let's keep testing in its rightful place - as a mere tool in the educational kit designed to educate and serve.

FOOTNOTES

[1]Romano, Octavio Ed. Quinto Sol Publications, Berkeley, California El Grito, Vol. II.

[2]Lambert, Wallace E. and Tucker, G. Richard. "The Home/School Language Switch Program in the St. Lambert Elementary School, Grades K-5."

[3]U.S. Commission on Civil Rights, "The Unfinished Education, Outcomes for Minorities in the Five Southwestern States". October, 1971.

[4]California State Department of Education. "Miller-Unruh Basic Reading Program". Annual Evaluation Report, 1969-70.

[5]Palomares, Uvaldo and Trujillo, Miguel P. First Quarterly Project Report on "Examination of Assessment Practices and Goals and the Development of a Pilot Intelligence Test for Chicano Children". O.E.O. Grant, Washington, D.C., October, 1971.

[6]Ibid.

[7]Sanchez, George I. "Bilingualism and Mental Measures: a word of caution". Chicanos Social and Psychological Perspectives, (1971). Natanial Wagner and Morsher J. Jaug, Eds.

[8]Jackson, Rex. Developing Criterion-Referenced Tests, Test Development Division, Educational Testing Service, Princeton, New Jersey. June, 1970. Distributed by ERIC Clearinghouse on Tests, Measurement and Evalution.

[9]Project Hacer Vida, Title VII Bilingual Education and Diagnostic Placement, Santa Ana Unified School District, Santa Ana, California.

TABLE OF CONTENTS

		Page
PREFACE		i
FOREWORD		v
PART 1	TESTING AND ASSESSMENT WORKSHOP	1
1.	Rationale for the Meeting	1
2.	Wechsler Intelligence Scale for Children (WISC)	5
3.	Comprehensive Tests of Basic Skills (CTBS)	10
4.	Cooperative Primary	15
5.	Lorge-Thorndike	19
6.	Culture Fair Intelligence Test	24
7.	Michigan Oral Production Test	28
8.	Peabody Picture Vocabulary Test	32
9.	Critique Guide	37
10.	Position Statement	41
11.	Resolutions	42
PART 2	PROFESSIONAL CRITIQUE OF THE NEW INTER-AMERICAN SERIES	43
PART 3	ABSTRACT: A SYSTEM FOR CRITERION-REFERENCED ASSESSMENT OF A BILINGUAL CURRICULUM	62
PART 4	SOME CAUTIONARY NOTES ON ATTEMPTING TO ADAPT IQ TESTS FOR USE WITH MINORITY CHILDREN AND A NEO-PIAGETIAN APPROACH TO INTELLECTUAL ASSESSMENT	65

PART 1 TESTING AND ASSESSMENT WORKSHOP

Rationale For The Meeting

In response to growing dissatisfaction among bilingual/bicultural educators, evaluators and psychologists with the continued use of standardized achievement and traditional IQ tests, BABEL held a Testing and Assessment Workshop in Berkeley, California on January 27-28, 1972. In attendance were approximately 150 bilingual psychologists and evaluators, educators working in bilingual/bicultural programs, and community representatives, from all over the Bay Area, Northern and Southern California, and Austin, San Antonio, Fort Worth and Crystal City, Texas.

The conference was planned with three specific objectives in mind. First, while people have repeatedly criticized existing tests being used in Bilingual Education Programs, few have actually documented where these tests penalize or harm bilingual/bicultural children. The first objective of the BABEL conference was to examine closely eight of these instruments and attempt to document harmful or inappropriate facets of them. The following tests, all used extensively in Bilingual Education Programs, were so discussed:

WISC (Weschler Intelligence Scale for Children)
CTBS (Comprehensive Tests of Basic Skills)
Cooperative Primary
Lorge-Thorndike
Inter-American Series--General Ability
Culture-Fair Intelligence Test
Michigan Oral Production Test
Peabody Picture Vocabulary Test

A second objective was to look at the Criterion Referenced models as a realistic alternative to traditional assessment. The third objective was to formulate and adopt a resolution(s) for consideration in Sacramento and elsewhere in the country.

The format of the conference was organized to facilitate the implementation of the above objectives. The conference opened on Thursday morning, January 27, 1972 with an informal coffee hour, followed by introductions and a welcome given by Dr. Rene Cardenas, Director of BABEL. Mrs. Olivia Martinez, Coordinator of Testing and Evaluation, gave a short background of the initial conception of the conference, and the responsibilities of those in attendance. The General Session was conducted by Dr. Ed DeAvila of the Multi-Lingual Assessment Program in Stockton, California. The text of his talk, "Testing Tonterias", is included here in this pamphlet. After a short break the entire group broke up into eight workshop sessions to evaluate the tests mentioned above. The workshop members were asked to examine, discuss and evaluate their test according to the following guidelines: vocabulary, illustrations, directions, lay-out design, cultural implications, translations, timing and scoring procedures, and norming of the test. A copy of the critique guidelines can be found in this pamphlet. (The workshop sessions lasted into the late afternoon). At the end of the sessions, each workshop member was asked to summarize his findings about the test in terms of the effectiveness and appropriateness of the test for use with bilingual/bicultural children in Bilingual

Education Programs. The members were also asked to complete and sign a position statement on the test, in which recommendations for the future use of the test were stated: continued use, modification, discontinue. A copy of the position statement appears in this pamphlet. Late in the afternoon a general session was held in which the findings and recommendations of each workshop were briefly summarized, discussed, and legal strategies considered.

The general session on Friday, January 28, 1972 dealt with alternatives to present standardized tests. Ed Apodaca, Director and Tomas Lopez, Evaluator of Project Hacer Vida spoke about "The Indio Criterion Referenced Model", and explained to those in attendance the formation, objective and use of the Criterion Referenced tests. Mr. Ben Soria, Director and Norm Nicolson, Evaluator reported on the Santa Ana Evaluation plan. It was felt that the Criterion Referenced Models along with attitudinal surveys, self-concept measures and other affective considerations should provide an appropriate and meaningful measure of program effectiveness. The rest of the morning and early afternoon was devoted to small grade level meetings in which the criterion referenced models, other alternatives, resolutions and position statements were discussed. Late in the afternoon another general session was held to draft group resolutions and discuss potential legal strategies.

The participants were also asked to fill out a form evaluating the various aspects of the two day conference. A

copy of this evaluation form is included in the pamphlet. The general opinion of the conference participants was very favorable.

BABEL is planning another Workshop to be held sometime during the school year 1972-73. This conference will concentrate on the Criterion-Referenced Models of assessment--how they are constructed and how they are to be used. BABEL also hopes to establish a means of training people in the uses and implementation of the Criterion-Referenced models in order that these models can be used in the trainee's school districts.

Wechsler Intelligence Scale for Children -- WISC

The Wechsler Intelligence Scale for Children grew out of the Wechsler-Bellevue Intelligence Scales used with adolescents and adults. The WISC may be used with children ages 5 through 15.

The WISC consists of 12 tests which are divided into two subgroups identified as Verbal and Performance. The tests of the scale are grouped as follows: Verbal; General Information, General Comprehension, Arithmetic, Similarities, Vocabulary and digit span. Performance; Picture Completion, Picture Arrangement, Block Design, Object Assembly, Coding or Mazes. Normally, 10 of these tests are given. Digit span and Mazes (or Coding) are considered supplementary tests to be added when time permits, or used as alternate tests. While the tests are identified as Verbal and Performance, and differ as these labels indicate, they each tap other factors, among them non-intellective ones, which produce other classifications or categories that are important in evaluating the individual's performance.

The theory underlying the WISC is that intelligence cannot be separated from the rest of the personality. An attempt is made, then to take into account the other factors which contribute to the total effective intelligence of the individual. The WISC renounces the concept of mental age as the basic measure of intelligence--I.Q.s are obtained by

comparing each subject's test performance exclusively with the scores earned by others in his own age group, rather than by comparing the performance with composite age groups. Also, no attempt has been made to define the social and clinical significance of any given IQ.

The group that evaluated the WISC was greatly concerned with the cultural orientation of the test. It was definitely felt that this test is not anywhere within the cultural reference of bilingual/bicultural children. The test is Anglo-culture oriented, and neither the illustrations nor the vocabulary can be generalized to other cultures. The consensus of the group was that this test is an unfair instrument to use in measuring the IQ of bilingual/bicultural children. When used with bilingual/bicultural children, the WISC measures acquired aculturation to mainstream middle class white culture, rather than I.Q.

In terms of directions and timing, the group felt that the WISC creates problems for the bilingual/bicultural child. The directions are too difficult in both the written and oral forms for these examinees. Many bilingual/bicultural children are unable to read the written directions because of their initial problems in learning to read English. Often, too, the oral directions are difficult because of the unfamiliar vocabulary that is used. The WISC is a timed test. The majority of the group was convinced that timed tests are not valid for testing bilingual/bicultural children, because they do not give an accurate picture of the actual abilities.

It was also felt that the WISC is too long, thus fatiguing the examinees, and again, not giving a true picture of ability.

The group was critical, too, of the lay-out design and illustrations used in the WISC. It was felt that there are few illustrations, and that there should be more available on the test. The lay-out design is also inadequate. There are too many items crowded onto each page, making the test confusing, especially for primary grade children.

The group was adamant about the fact that this test was not developed for testing bilingual/bicultural children. Considering this fact, the group felt that translating this test directly into Spanish would not make it more valid. The group decided that a translation of the WISC would have to take several things into account. First, a translation would have to be correlated with classroom instruction and activities in bilingual education programs, in order to give the test validity. Secondly, any translated version of the WISC would have to consider the many regional variables in written and spoken Spanish in the United States. These variables would have to be included in the test, and accepted as correct responses where applicable.

The group felt that the result of the WISC are absolutely confusing and meaningless for bilingual/bicultural children. The group was concerned about the fact that the results of

the WISC would label bilingual/bicultural children and negatively affect teachers' attitudes toward students. The majority of the group seemed to feel that the WISC could possibly be used only as a diagnostic test, but that it is totally invalid as an intelligence test when used with bilingual/bicultural children.

It should be noted that there is a movement in the California State Department of Education to initiate a project to renorm the WISC for the bilingual/bicultural population of the state. The evaluation group was definitely against the renorming of the WISC for the following reasons:

A. There is a Spanish version of the WISC already developed in Puerto Rico which is not desirable because it does not include regional variations in the Spanish language.

B. The researcher presently involved in the renorming project is not bilingual/bicultural.

C. Research shows that bilingual children generally do not benefit from taking a Spanish version of an I.Q. test.

D. The population that would be normed is linguistically very diverse in Spanish, which would make the renorming of this test difficult, and the results, at best, vague.

E. The group rejects the use of I.Q. as a solitary measure of the intelligence of bilingual children.

F. There is a need for the development of criterion reference measures to determine the abilities of bilingual/bicultural children.

In conclusion, the group seemed to feel that the WISC can not effectively evaluate either the success or weakness of a bilingual program, the potential and I.Q. of bilingual/bicultural children, or what these children learn in bilingual education classes. It was concluded that the WISC does not reveal to the classroom teacher how she might improve her teaching. The group was concerned with the fact that taking this test might definitely be harmful to the bilingual/bicultural child, unless the test was used for diagnostic purposes. The majority of this group seemed to feel that the WISC could not be effectively modified for use in bilingual education programs, and that new instruments should be developed to replace the WISC.

The group recommended that the WISC be discontinued as an evaluative tool for bilingual/bicultural populations, but that its use be continued for individual diagnostic purposes on special children with certain learning difficulties. The group suggested "...that an organized group of bilingual/bicultural psychologists (i.e. through CASPP) recommend to the State Department of Education or to the State Legislature or to whomever can effect change in each state, that any existing version of WISC be discontinued as a measure of intelligence when used with bilingual/bicultural children."

The Comprehensive Tests of Basic Skills -- CTBS

The CTBS are a series (of batteries) of 10 tests in four basic skills areas: reading, language, arithmetic and study skills. There are four levels of the CTBS in the series, designed as follows: Level 1 for grades 2.5-4.9; level 2 for grades 4.0-6.9; level 3 for grades 6.0-8.9; and level 4 for grades 8.0-12.9. The overlapping levels provide the user with a choice of level for use in Grades 4,6 and 8.

These tests were designed to measure the extent to which the individual student has developed the capabilities and learned the skills which are pre-requisite to the study of specific academic disciplines. The emphasis in this series is on the measurement of (the grasp of) broad concepts and abstractions developed by all curriculums, and on facility in such skills as classifying, manipulating, translating and interpreting, which are needed in the effective use of language and number. These tests are not like basic achievement tests in that they are not affected by the content material used to teach students. Performance is affected by the grade level at which topics are introduced into the curriculum and by the development of the necessary capabilities to perform the tasks.

The test items in the CTBS for the four skills mentioned above generally measure the following: the ability to recognize and/or apply techniques, including performing

fundamental operations; the ability to translate or convert concepts from one kind of language (verbal or symbolic), to another; the ability to comprehend concepts and their interrelationships; the ability to extend interpretation beyond the stated information.

In evaluating the CTBS for use with bilingual children and/or in a bilingual program, there were several general considerations that concerned the group of evaluators.

Of primary concern was the fact that the CTBS is oriented toward the Anglo culture, Anglo study skills and school situations and obviously, Anglo use of the English language. This orientation might make the CTBS fairly effective when used with white middle class Anglo students. The same orientation renders the CTBS highly ineffective and inappropriate when used to evaluate the abilities of bicultural or bilingual students. There is little in the CTBS that bicultural/bilingual children can relate to and it is significant that tests like the CTBS do not have multi-cultural considerations, so as to be appropriate for those who must take the test. Thus, it was felt that the CTBS is being used presently in the state of California not because of its effectiveness, but for two very different reasons. First, it is a state mandated test, that is, it is designated for use in the public schools of California by the State Department of Education. Secondly, results of the CTBS show bicultural/bilingual students functioning far below grade

level, and these results are used as a vehicle for obtaining state and federal financial aid for various school districts.

The group was also concerned about the directions used in administering the CTBS. It was felt that the bilingual/bicultural child could have difficulty understanding the directions of this test because of a possible limitation in knowledge of English, and his problems with written English. It was decided that translating the directions into written Spanish would not be beneficial to the bilingual child because of the fact that many Spanish-speaking children in the United States are illiterate in Spanish.

It would be invalid to translate the existing CTBS into standard Spanish not only because of the illiteracy problem, but also because of the regional differences both in Spanish language and culture, make it very unlikely that the test could be normed for a general area or region of the country. The possibility of a national standardized test for bilingual children was discussed in these terms, and was rejected by the group of evaluators.

The evaluators considered the lay-out design of the CTBS very confusing for the examinees. The pictures, questions and phrases are poorly spaced on the pages of the test. They are cramped together and present an unorganized and indiscriminate set of stimuli for the examinees.

The group rejected the concept of the timed test. The timed test adds a great deal of tension and pressure to the testing situation, and thus tends to give an imprecise picture

of the examinees' abilities. It was also felt that the CTBS is too long and tends to fatigue the examinees. The fatigue factor can affect the results of the test.

Another consideration that the group was concerned with was ~~the scoring~~ procedures of the CTBS. They felt that there is little correlation between the test scores and actual classroom behavior. There seem to be many instances where the examinees do poorly on the test but are progressing fairly well in the classroom. It was also felt that any type of score on a test such as the CTBS is dangerous in that it tends to be misused by classroom teachers in labeling students and student potential, and thus creating a particular bias in teacher attitudes toward students.

The group that evaluated the CTBS came to several conclusions about the appropriateness of this test for use with bilingual/bicultural children. It was decided that the CTBS effectively evaluates neither the potential of the bilingual/bicultural child, nor what a child has learned in bilingual education classes, nor the successes or weaknesses of bilingual education programs. It was felt that the CTBS is also ineffective in revealing to the classroom teacher how she may improve her teaching. Moreover and more important, it was decided that taking this test is of negative value to the bilingual/bicultural child, and may very possibly be harmful to him. The majority of the group felt that there was a possibility that this test could be modified for use in a bilingual education program, but that the success

of such modification was doubtful.

The evaluation group recommended that the CTBS be discontinued as an evaluative tool for bilingual/bicultural populations. The group felt that, "A new instrument should be developed which takes the child's cultural reference and language capabilities into consideration. The CTBS scores are not only invalid, but in many instances are detrimental to the self-concept of bilingual/bicultural students."

Cooperative Primary Reading Tests

The Cooperative Primary Tests are carefully constructed and standardized general achievement tests. As such, they may be expected to serve a wide variety of educational and administrative purposes. One of the major purposes of the test series supposedly is to provide teachers with measures of children's concepts and skills that closely relate to their work in the classroom. Identify other forms, i.e., math.

The skills being tested on the cooperative Primary Reading Tests are representative of three categories: Comprehension, which asks for identification of an illustrative instance and/or identification of an associated object or instance; Extraction, which asks for the extraction of an element or elements, the extraction of an element in order, or the identification of an omission; and Interpretation, Evaluation and Inference. There are no time limits. The children are allowed as "reasonable" amount of time to finish the test.

The Cooperative Primary Reading Tests were normed in April of 1966 in four regions of the United States. Approximately 1700 public school children at 170-176 schools made up the sample at each grade level. The data gathered from these administrations were used to develop scaled scores, percentile ranks, stanines and grade equivalents. The Cooperative Primary Reading Tests yield one score: the total number of correct responses. The most widely used

means of interpreting this raw score for these tests is percentile ranks. In some cases test scores may provide the teacher with important cluee about the achievement of a child. "In most cases, however, test scores will serve primarily as verification of her judgement." (manual p.8)

There was general agreement among the members of the group that evaluated the Cooperative Primary Tests, that these tests have some value when used with the intended population-similar to the norming population-middle class, monocultural, English-speaking children. The group unanimously felt that these tests do not have validity when used with bilingual/bicultural children. In coming to these conclusions, the group dealt with several basic considerations.

First, the group was concerned with the vocabulary items used on the Cooperative Primary Test. It is much too advanced, not only for a heterogeneous group of examinees that includes bilingual/bicultural children, but most probably for any child taking the test. It was felt that many of the items are difficult, and inappropriate, because they are regionally oriented. Words like "mitten" and "snowman" are for many pupils, inappropriate items when used in certain regions of the country. Many of the evaluators also felt that the vocabulary used in the directions is difficult, and particularly so for bilingual/bicultural children. A major criticism, too, was that the directions are too lengthy, almost to the point of becoming unclear.

Of particular concern to this group of evaluators was

the general lay-out design and visual presentation used in the Cooperative Primary. It was felt that the lay-out is definitely too crowded, both picture and vocabulary items are packed together on each line and each page. The total effect is very distracting and confusing for the children. One evaluator made the comment that children are not taught to read in the manner in which the words are positioned on this test. The group was critical, too of the illustrations used. While the quality of the pictures themselves is acceptable, it was felt that some of the illustrations a misleading, for example the snowman, snail, mitten and elephant, do not accurately depict the desired responses. There was concern among some of the evaluators, too, that several of the pictures are not easily identifiable to bilingual/bicultural children. "The children have never been exposed to many of the pictures given." An additional suggestion was made to darken the lines of the arrows in which the cues are written.

Although the Cooperative Primary is not a timed test, the group seemed to feel that bilingual/bicultural children would not have sufficient time to finish the test. There was also the feeling that the competitive factor is significant, "The administration of a test that requires competition is not usually beneficial to bilingual/bicultural children." The group also felt that the length tends to frustrate children.

The group was adamant in its opinion that the results of the Cooperative Primary Tests are unclear and meaningless

for bilingual/bicultural children. In this case, it was felt that the results not only do not help a teacher to understand and help students, but are also misleading and harmful. "Tests such as this are the basis for pegging minority children and placing them in MR classes." "It is mandated by the state of California, an absurd requirement." "A crime to use it."

The group came to some final conclusions about the Cooperative Primary Reading Tests and their appropriateness for use with bilingual/bicultural children. It was felt that this achievement test evaluates neither the potential of bilingual/bicultural children, nor what a child might learn in bilingual education classes. This test could not evaluate the successes and weaknesses of a bilingual education program either, nor does its use necessarily reveal to the teacher how she might improve her teaching. There was a definite consensus among the group that this test is of negative value and harmful if given to bilingual/bicultural children. The group felt that there was no realistic possibility of modifying this test for use in bilingual education programs. "Modification would merely give a semblance of validity to an invalid instrument."

The evaluation group recommended that the Cooperative Primary Reading Tests be discontinued as an evaluative tool for bilingual/bicultural populations. Many group members also felt strongly that this test should not be used under any circumstances for bilingual/bicultural children.

Lorge-Thorndike Intelligence Tests

The Lorge-Thorndike Intelligence Tests are a series of tests of abstract intelligence. That is, they test the ability to work with ideas and the relationships among ideas. The tests are based upon the premise that most abstract ideas with which school children or working adults deal are expressed in verbal symbols. Thus, verbal symbols are the appropriate medium for testing abstract intelligence. These tests take into account the fact that for some - "the young, the poorly educated, or the poor reader" - printed words are an inadequate measure of abilities. Consequently, a parallel set of nonverbal tests is provided.

There are two batteries of tests. The Primary Battery is used with subjects in kindergarten through third grade, and consists of two levels. There are three subtests for each level, Oral Vocabulary, Cross-Out and Pairing. Each requires less than ten minutes for administration. However, the test is untimed and the administrator adjusts the pace to the students. The Multi-Level Battery tests subjects in third grade through college and has eight levels. The term "multi-level" indicates that there is a graded series of items divided into eight different but overlapping scales for use within the grade range. There is a separate series of items for each grade in the lower end of the overall grade range and a separate series of items for each pair of grades in the upper part of the grade range.

The Verbal series of the multi-level tests is made up of five subtests which use only vocabulary items: vocabulary, verbal classification, sentence completion, arithmetic reasoning and verbal analogy. The Nonverbal series uses items which are either pictorial or numerical. It contains three subtests involving picture classification, pictorial analogy, and numerical relationships. The working time for the Verbal series is 35 minutes and for the Nonverbal series is 27 minutes. It is suggested that both series of tests be used for the appraisal of children in schools.

The Lorge-Thorndike Intelligence Tests were evaluated by a group that was concerned with several general considerations.

Of primary concern to the evaluators was the fact that the Lorge-Thorndike is currently being used in the state of California, as a state mandated test. It was felt by the group that the test is ineffective, and has been indiscriminately designated for use in the public schools. Secondly, the group felt that the Lorge-Thorndike is being used at the expense of bilingual/bicultural children. The results usually show bilingual/bicultural students functioning far below grade level and these results are used as a vehicle for obtaining state and federal financial aid for various school districts.

The evaluators considered the Lorge-Thorndike to be culturally biased. The test is definitely oriented toward the Anglo culture, and offers little that is within the bilingual/bicultural student's cultural reference. This

renders the Lorge-Thorndike highly ineffective and inappropriate for measuring the I.Q. of these examinees. The group felt that any test like the Lorge-Thorndike must have multi-cultural implications, so as to make it appropriate for those who must take the test.

The group felt that the directions used in administering the Lorge-Thorndike Tests are too difficult for the bilingual/bicultural child, because of his possibly limited knowledge of English and problems with written English. It was decided that translating the directions into written Spanish would not be beneficial to the bilingual child because of the fact that many Spanish-speaking children in the United States are illiterate in Spanish. Direct translation of the Lorge-Thorndike into Spanish would solve nothing. It would be invalid to translate the existing Lorge-Thorndike tests into standard Spanish, not only because of the illiteracy problem, but also because of the regional differences in the Spanish spoken by children who would take such a test. Regional variables both in Spanish language and culture make it very unlikely that the test could be normed in Spanish for a general area or region of the country.

The group also rejected the concept of the timed test. The time competition on the Lorge-Thorndike definitely puts the bilingual/bicultural child at a disadvantage. The timed test adds tension to the testing situation, and tends to give an imprecise picture of the examinees' abilities. It was also felt that the Lorge-Thorndike is too long and tends to

fatigue the examinees.

The group was also concerned with the scoring procedures of the Lorge-Thorndike tests. They felt that the results are meaningless for bilingual/bicultural students. They also seemed to think that there is little correlation between the test scores and actual classroom behavior. There seem to be many instances where the examinees do poorly on the test but are progressing fairly well in the classroom. It was also felt that any type of score on a test such as the Lorge-Thorndike is dangerous in that it tends to be misused by classroom teachers in labeling students and student potential, thus creating a particular bias in teacher attitudes toward students.

Some comments made by individual evaluators in the group are interesting and valid:

"The Lorge-Thorndike is basically a reading test--therefore, if a child can't read, he is unable to take the test. The test is much too difficult for the bilingual child...".

"I can see no positive values to the child taking the test. It should not be used for tracking children."

"This test is a reading test. Not only is this test ineffective for bilingual children, but for any child that cannot read."

The group came to the following conclusions: The Lorge-Thorndike does not effectively evaluate the success or weakness of a bilingual program. The test does not measure the potential of bilingual/bicultural children, nor their I.Q.,

nor what they learn in bilingual education classes. The group felt that the test does not reveal to the teacher how she may improve her teaching, but allows her to label children based upon the test scores. It was the general consensus of the group that this test is of no positive value to the child, but, on the contrary, tends to be harmful. The majority of the group felt that modification is not the answer in dealing with the Lorge-Thorndike tests.

The group recommended unanimously that the Lorge-Thorndike tests be discontinued as an evaluative tool for bilingual/bicultural populations. There were also several recommendations that this test not be used under any circumstances for bilingual/bicultural children.

The Culture Fair Intelligence Test

The Culture Fair Test is an intelligence test which claims to measure intellectual capacity or potential; that is, it measures the child's capacity to learn in the future, rather than his already learned scholastic skills. The test is perceptual and nonverbal and thus claims to give fair predictions of future potential when used with children from diverse homes and cultural backgrounds. Experiential differences, i.e. opportunities in the years preceding the test have been shown to have a powerful effect on the outcome of most intelligence tests now in use, but the Culture Fair Test claims to have been used with equal success on children from various backgrounds.

The Culture Fair Tests consists of three scales: Scale 1 for ages 4 through 8; Scale 2 for ages 8 through 14, (grades 3 through 9); and Scale 3, designed to discriminate in the upper ranges of intelligence for young adults and adults. Scale 2 consists of two parallel forms, A and B, each totaling 46 items arranged in four subtests and covering 12 and 1/2 minutes of test time. The two forms permit a rest pause half way in the testing, and interruption of the test for completion on another day, or an I.Q. assessment based on a single form if time is short.

The Culture Fair Test has four subtests in each of the two forms. The content is figural and geometric, and each subtest involves a different kind of test question: Series,

Classification, Matrices, Conditions, (topology). It is possible to use the same form or forms for retest or for testing at yearly intervals.

The group that evaluated the Culture Fair Intelligence Test was primarily concerned with the fact that this test is not a valid measure of I.Q. for bilingual/bicultural children. It was felt that the Culture Fair provides some measure of abstract reasoning and of spatial perception, and as such, the test does have some validity. However, the group was critical of various aspects of the test in its present form.

First, the group was concerned with the English vocabulary used in the directions of the test. It was felt that the vocabulary is very difficult for bilingual/bicultural children. The directions are lengthy and ambiguous, and certainly not appropriate for the age groups being tested. The directions have been translated into standard Spanish, but the group found the translation to be unsatisfactory. Because the test is translated into Standard Spanish, no consideration is given to regional differences in the language. The Spanish used here is very sophisticated and literary, thus possibly inappropriate for use with some bilingual/bicultural children in the U.S. "The words used in the Spanish directions are terms that not many Mexican-American children can understand."

Because the test is nonverbal, and is composed of various geometric figures, it was felt that the quality of the lay-out design and the illustrations is of special importance. The group was critical of the lay-out of the test. The items

are very crowded..."there is an overwhelming amount of clutter...". It was decided that the test definitely needs a new lay-out, in which items are better spaced on the pages. There was some concern as to the appeal of these geometric figures for small children...Are the figures interesting to the children, or are they frustrating?

This test is designed to be culture-free, or culture-fair, and thus, fair in its measurement of the I.Q. of all children. However, the evaluators were concerned with one thing which they felt made the test culturally biased. The test is timed, and makes competition an important factor in taking this test. The evaluators referred to the time-competition factor as the "American Approach". It was felt that this factor is often culturally alien to bilingual/bicultural children, that it creates tension and frustration, and handicaps them in the testing situation. The group mentioned, too, that the time allotted was definitely inappropriate for those taking the test. Distinguishing between geometric figures is a very precise and demanding activity and demands more time for the testees than is allotted. The group also concluded that the length of the test is definitely fatiguing to the children considering the types of activity demanded here.

In conclusion, the evaluators decided that the Culture Fair Intelligence Test in its present form, and with its present purpose, does not effectively evaluate either the strengths and weaknesses of a bilingual program, or the I.Q. of bilingual/bicultural children. It was felt that it cannot

effectively evaluate what a child has learned in bilingual education classes unless his curriculum has focused on spatial perception and abstract thinking. "If this type of test is to be administered to children, then they should be taught perceptive and abstract reasoning concepts in the classroom...".

It was felt that the test is positive in that it may reveal to a teacher how she can improve her teaching, in terms of focusing on perceptual concepts and abstract reasoning. The group seemed to feel that the test may have positive value for the children as well if it is used correctly. What looms as a negative factor for the children is that time competition factor of the Culture Fair test. In general, it was felt that this test can and should be modified, by changing the focus of the test from I.Q. to diagnosis. "It can be used as a diagnostic tool in order to help teachers find areas of weakness, rather than as an instrument to stratify some children."

The group recommended that the Culture Fair Intelligence Test be discontinued as a measure of the I.Q. for bilingual/bicultural children. They recommended that with modification this test could be used for diagnostic purposes, and as a measure of spatial perception and abstract reasoning with individual children or small groups. "This insturment should not be labelled as an I.Q. test, but as a test of perception and abstract reasoning. This instrument could have great implications for the development of curriculum and teacher training, and as a diagnostic tool."

Michigan Oral Language Productive Test--Structured Response.

The Michigan Oral Language Productive Test is based uon upon the Dade County Test of Language Development--the original test has been revised and enlarged to 43 items. The purpose of the test is to assess a child's ability to produce standard grammatical and phonological features when he speaks English.

The method used in administering this oral language test is the following: The child is shown three pictures which form a story. He is given a Stimulus concerning one of the pictures. The stimulus is structured so that the child will give a response containing a particular feature of grammar or pronunciation. For example:

Question 5--Stimulus
Past Participle

Stimulus--(point to boy in picture) (Child's name). Ask the boy if he always goes to this river to fish.

Have you always....

It is stressed that the standard stimulus (given in the test manual) always be given as it is written, in order that there be a cue that evokes the desired response. It is also very important when using the Michigan Oral Language Productive Test to set the child at ease before beginning the test (provide a verbal warm-up period), and to praise the child when he speaks, with moderately positive comments such as fine or You're giving me lots of answers.

There are 43 items on the test, which should take approximately 15 minutes to give. These items represent 11 categories of grammatical and phonological features: uses of be; uses of have, comparison, uses of do, double negative, past tense, past participle, plural, possessive, pronunciation, subject-verb agreement. The scoring sheet provides for various alternatives to the standard desired responses. From this sheet, category percentages for the eleven categories can be determined.

It is stressed that the value of the structured response test is its ability to give the teacher a quick overview of her student's language needs. The more efficient the curriculum is in meeting the students' language needs, the more quickly the overview is likely to change.

The Michigan Oral Language Productive Test was evaluated briefly, and the group seemed to find some value in using the test with bilingual/bicultural children, although it does not evaluate the potential or the I.Q. of a bilingual/bicultural child. It was felt that the test does effectively evaluate what a child has learned in the English or ESL component of a bilingual education program. Similarly, the test does point out the strengths and weaknesses of the English language or ESL curriculum of a bilingual program, by pointing out the particular language strengths and needs of the children. And, too, the Michigan Oral Language Productive Test can show a teacher where she needs to improve her language teaching, although it does not reveal exactly how she may improve it.

In dealing with the question of the positive or negative

value of the test for the bilingual/bicultural child, the group placed responsibility with the test administrator, the teacher. Most of the group felt that taking the test can be a threatening experience for the child if the administrator is unable to put the child at ease and make him feel that his responses are successful. It was felt that the test can have positive value if the child is skillfully praised for his responses and can take the test in a comfortable and relaxed environment.

It was suggested that the Michigan Oral Productive Test can be successfully modified to meet more of the evaluative needs of bilingual education programs. One suggestion was to change the stimuli, or the order of the stimuli in order to bring the test closer to actual classroom curriculum. Another suggestion was to make the test more bicultural by making the test pictures relevant to the culture of the examinees, rather than settling for "Anglo prototype" pictures. A final concern was that this test "needs guidelines to determine what sequence learning should take.... the test assumes teacher objectivity and ability to improvise beyond the capabilities of some teachers."

The majority of the group recommended that the test be continued for use with bilingual/bicultural populations, or for individual diagnostic purposes on special children. Most of the members had reservations about complete and open endorsement of the Michigan Oral Language Test: "until something better is developed";"only with modification;

"to help develop ESL lessons for a particular class only"; "for testing ESL only"; "to measure the extent to which a child speaks English only".

The Peabody Picture Vocabulary Test

The Peabody Picture Vocabulary Test is designed to give an estimate of a child's verbal intelligence, by measuring his hearing vocabulary. The test consists of 150 plates preceded by three example plates. The examiner asks the examinee to identify various vocabulary items by pointing to the picture on each plate that best tells the meaning of each item. The test is untimed, but takes usually 10 to 15 minutes.

There are two forms to the test, A and B, each consisting of 150 plates. The plates are arranged in empirically-determined order of difficulty. A fairly even number of plates are placed at each age level with a somewhat heavy concentration at the pre-school levels. The four vocabulary items used to make up each plate were selected based upon the following criteria: all four words were found to be at the same difficulty level; all four words demonstrated good linear growth curves; words were used where no sex differences were found to exist; primarily singular and collective nouns, some gerunds, and a few adjectives and adverbs were used; words were omitted which seemed to be biased culturally, regionally, and racially, as were dated words, plurals, double words, scientific terms, etc.

The illustrations were selected based upon the following criteria: equal size, intensity, and appeal; and appropriateness to the age level of the subjects most likely to

view the plate.

Besides being effective with average subjects, the PPVT has special value with certain other groups of subjects. Since subjects are not required to read, the test is used with non-readers and remedial reading subjects. Because the responses are non-oral, the test is appropriate for children with speech impediments, and for certain autistic or withdrawn children. The test has also been used with handicapped and perceptually impaired subjects. "The scale may be given to any English speaking resident of the United States between 2 years 6 months and 18 years who is able to hear words, see the drawings, and has the facility to indicate "yes" and "no" in a manner which communicates." (p. 25-manual)

The group that evaluated the Peabody Picture Vocabulary Test felt that in general this test was one of the best instruments that is currently available for measuring the capabilities of bilingual/bicultural children. It was the consensus of the group that the basic structure of the test is good, that the directions are clear and appropriate for the children being tested, that the lay-out design is good, and that the untimed nature of the test is a positive quality. There were, however, several negative considerations that the group discussed in relation to the PPVT.

First, it was felt that the illustrations used in the PPVT are very flat and not as appealing as they could be. The group suggested that the pictures should be in color, and

should be larger for use with the very young examinees. The group found that some of the items were ambiguous, for example, item 22-which contains pictures of what appears to be a boat on a river, some vegetables, a rosebush and two mountains. Finally, and very importantly, it was felt that many of the illustrations have cultural references that are not easily identifiable to the bilingual/bicultural child (items 31, 44, 57, 58, 62, 64, 27, 69, 70, 71). In regard to cultural implications the group seemed to feel that many of the items in the PPVT are not fair to bilingual/bicultural children. Few of the items are reflective of the cultures of bilingual children. It was felt that as the test progresses, the items definitely become more culturally complicated, and move further away from what the bilingual/bicultural child can relate to culturally. For example, item 64 contains the pictures of a fencer, prim lady with pencil and paper, an old woman giving a speech at a podium and a chef standing at the stove, which are all strange items for some bilingual/bicultural children. It is interesting to note that this test was normed on a population of white Anglo children living around Nashville, Tennessee.

The group was also concerned with the vocabulary used on the PPVT. Not only are many of the pictures on the test oriented toward anglo culture, but the desired vocabulary responses also reflect this orientation. The group seemed to feel that the first ten items and vocabulary responses are applicable to the bicultural/bilingual child, but that after

item ten, the vocabulary becomes increasing more difficult and more unrelated to the child's language experiences. They cite as an example, item 22, which asks for the response "bush". Item 67 asks for the response "stadium". Item 70 asks for "stunt". Item 71 asks for the word "meringue", and 72 asks for "appliance". The group proposed an evaluative study of words that are familiar and relevant to the bilingual/bicultural child.

In discussing the possibilities of translating this test into Spanish, the problem of regionalism was of great concern. Spanish vocabulary items definitely vary depending upon the region of the United States. For example, the group cited several possibilities in Spanish for the word truck; troca, camioneta, camion; for the word car; carro, auto, automovil; for the word baby; babito, nino nene; for the word teacher; maestra, profesora. The group concluded that an extensive study of the regional differences in Spanish should be made. It was felt that after such a study, regional differences should be taken into account in the PPVT, and that these different forms should be included in the test and be accepted as correct responses in each particular region.

In conclusion, the group decided that the PPVT does not, in its present form, effectively evaluate the success and weakness of a bilingual program, the potential of bilingual/bicultural children, nor what these children learn in bilingual education classes. It was felt that the relative value of this test for the examinee depends very much on how the results are

used. It was stressed that this test should not be used to measure I.Q., but more as a measure of vocabulary comprehension and growth.

The evaluation group for the PPVT recommended that the test be modified. They felt that the basic structure of the test was a good one. An attempt to modify the test should concentrate both on changing the anglo cultural orientation of the illustrations and the vocabulary and on the problems of regionalism in Spanish vocabulary. "The PPVT has good possibilities for development as an evaluative tool for bilingual programs. The test, in a modified form, could be used as a means of measuring the overall success of a bilingual program. It could also be used to measure the progress of children in a bilingual education program over a year's time."

BABEL TESTING AND ASSESSMENT WORKSHOP

CRITIQUE GUIDELINES

I. VOCABULARY

a. Is the content appropriate? i.e., do the words used adequately reflect those of the age group tested? _____ _____

b. Degree of difficulty. Are the words used too advanced or too easy for the test level? _____ _____

c. Visual presentation, positioning. Are the words arranged in an easy to read fashion? _____ _____

d. Other _____ _____

II. ILLUSTRATIONS

a. Are they ambiguous? i.e., can you tell easily what each drawing is supposed to be? _____ _____

b. Are the pictures of good quality? i.e. appealing to children? _____ _____

c. Cultural implications, do they depict items naturally and easily identifiable with Chicano or Asian cultures? _____ _____

d. Other _____ _____

III. DIRECTIONS

a. Are they clear? _____

b. Are the words used to instruct the children appropriate for their age? _____ _____

c. Are they very lengthy so that the point becomes unclear? _____

d. Other _____

CRITIQUE GUIDE - PAGE TWO

IV. LAY-OUT DESIGN

a. Position of items - are they items placed so that they bias other items? Are they positioned sequentially or randomly?

b. Visual Effect - is the overall impact an appealing one? Are they spaced far enough apart or are the items crowded?

c. Does one part of the test distract from another?

d. Other?

V. CULTURAL IMPLICATIONS

a. Are the items reflective of bilingual cultures?

b. Can the illustrations and vocabulary be generalized to other cultures?

c. Are the items "fair" to children who are bilingual/ bicultural?

d. Other?

VI. TRANSLATIONS

a. Are they correct?

b. Is the vocabulary used appropriate for children?

c. Are regional differences in language a factor?

d. Other?

CRITIQUE GUIDE - PG. THREE

VII. TIMED TESTS

a. How significant is the competitive factor? _____

b. Is the time allowed appropriate for children? _____

c. Other? _____

VIII. SCORING PROCEDURES

a. Are the results meaningful? _____

b. Are the results clear? _____

c. Do the scores/results help the teacher to understand and help her students? _____

d. Other? _____

IX. OTHER CONSIDERATIONS

a. Length of test by subsection and total, is it fatiguing to children? _____

b. What population was the test normed on? _____

c. How large was the norming population? _____

d. Does the test appear to be used the way in which it was intended by the author? _____

e. Other? _____

CRITIQUE GUIDE - PAGE FOUR

X. SUMMARY

a. Does this test effectively evaluate the success of a bilingual program? _____

b. Does the test effectively evaluate the potential of bilingual/bicultural children? _____

c. Does the test effectively evaluate what a child has learned in bilingual education classes? _____

d. Does the test effectively evaluate a bilingual child's I.Q.? _____

e. Does the test effectively evaluate the weaknesses of a bilingual program? _____

f. Does the test reveal to the teacher how she may improve her teaching? _____

g. Is taking this test of positive value to the child? _____

h. Is taking this test of negative value or harmful to the child? _____

i. Other? _____

BABEL TESTING & ASSESSMENT WORKSHOP

POSITION STATEMENT

I, have reviewed the _____ test
(please note level & form)
in terms of its appropriateness for use in evaluating bilingual/bicultural children and bilingual/bicultural programs.

_____ I endorse its continued use for bilingual populations.

_____ I endorse its continued use only for individual diagnostic purposes on special children with certain learning difficulties.

_____ I cannot give an opinion on this instrument (explanation attached).

_____ I urge that this instrument be discontinued as an evaluative tool for bilingual/bicultural populations.

_____ This test should not be used under any circumstances for bilingual/bicultural children.

NAME _____

POSITION _____

PROJECT/DISTRICT _____

DATE _____

RESOLUTIONS
(Drafted January 28, 1972)

1. Testing of children whose language is other than standard English with instruments that were developed for the user of standard English violates the norms and standardization of those instruments and therefore raises serious questions as to the results obtained. We, therefore, take the position that such users of these instruments with children whose language is othe han standard English is invalid.

2. Sufficient evidence now exists to direct us to the development of Criterion Referenced Assessment systems as a means of improving educational programs accountability for learning activities. It is imperative that these evaluation processes be correlated with local performance objectives.

3. The development of valid test instruments for bilingual/ and/or bilcutral children must be directed by bilingual and/or bicultural qualified personnel in the education field or similar fields; otherwise, the test instruments will not reflect the particular values, skills, etc. of the ethnic or cultural group being tested.

4. Whereas currently used standardized tests do not measure the potential and ability of California bilingual or bicultural children, and whereas these tests are being used if they do so measure, and they are relied upon to counsel, place and track these children, this body hereby resolves that such use of standardized tests should be immediately discontinued.

PART II

A CRITICAL REVIEW OF THE NEW INTER-AMERICAN SERIES

Prepared by Barbara Havassy, Ph.D., Consultant

For Multilingual Assessment Program

(Joe R. Ulibarri - Project Director)

The project presented or reported herein was performed pursuant to a Grant from the U.S. Office of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

The New Inter-American Series: Tests of General Ability & Tests of Reading

Author: Herschel T. Manuel

Publisher: Guidance Testing Associates

Author's Purpose: The Series consists of two types of tests: tests of general ability and tests of reading.

(1) Tests of General Ability: "...designed to provide an estimate of the ability to do academic work in general... The verbal materials test the understanding of written language and the ability to recognize relationships among concepts expressed by words. The nonverbal materials also present problems of relationship among concepts, but in these exercises the problems are expressed by pictures or drawings with only initial verbal directions. In the numerical materials, the ability to think quantitatively is tested by exercises in arithmetic computation and by exercises in arithmetic reasoning... The tests provide an estimate of abilities which cut across different fields of study." Not intended as a measure of general intelligence.

(2) Tests of Reading: These tests not only measure achievement in reading, but form a basis "for estimating ability to do school work in other areas in which the ability to read is related to achievement."

Description of Series: All tests at all levels available in English or Spanish editions.

A. Tests of General Ability

1. Pre-School Level. Individually administered in 2 periods with picture stimulus cards. Requires no oral response. Verbal-numerical: 40 items, non-verbal: 40 items.

2. Level 1 Pretest. Grades K-1. 4 page practice test to prepare for actual Level 1 test.

3. Level 1. Grades K-1. 80 items. A 'readiness' test. Administration in small groups (8-12 children) recommended. Consists of:

    1. Verbal-Numerical subtest, 40 items composed of oral vocabulary (part 1) and number (part 2) items.

    2. Non-Verbal subtest, 40 items composed of association (part 3) and classification (part 4) items.

4. Level 1. Abbreviated edition. Grades K-1. 64 items (fewer items in each area than the long form).

5. Level 2. Grades 2-3. 100 items. Consists of:

    1. Verbal-Numerical subtest, 60 items.

    2. Non-Verbal subtest (classification and analogies), 40 items.

6. Level 3. Grades 4-6. 150 items, 52 min. Consists of·

    1. Verbal subtest, 50 items. Composed of sentence completion (part 1) and word relations (part 4) 18 min.

    2. Non-Verbal subtest, 50 items. Composed of figure analogies (part 2) and figure classification (part 5) 16 min.

    3. Numerical subtest, 50 items. Composed of computation (part 3) and number series (part 6). 18 min.

7. Level 4. Grades 7-9. 150 items, 52 min. Same format as Level 3.

8. Level 5-Advanced. Grades 10-13. 150 items, 52 min. Same format as Level 3.

B. Reading Tests

1. Level 1. Grade 1. 80 items, 18 min.

    1. Part 1 Vocabulary 40 items, 8 min.

    2. Part 2 Comprehension 40 items, 10 min.

2. Level 2. Grades 2.5-3.0. 110 items, 23 min.

    1. Part 1 Level of Comprehension 40 items, 10 min.

    2. Part 2 Speed of Comprehension 30 items, 5 min.

    3. Part 3 Vocabulary 40 items, 8 min.

3. Level 3. Grades 4-6. 125 items, 41 min.

    1. Part 1 Vocabulary 45 items, 10 min.

    2. Part 2 Speed of Comprehensic , 30 items, 6 min.

    3. Part 3 Level of Comprehension, 50 items, 25 min.

4. Level 4. Grades 7-9. 125 items, 41 min. Same format as Level 3.

5. Level 5. Grades 10-13. 125 items, 41 min. Same format as Level 3.

Format:

Tests of General Ability: Levels 1 through 2 consist of pictorial items where the child marks in the test book the picture which is his answer to the question. For certain items (oral vocabulary and numerical) the teacher reads the question. For other items (classification and association) the question is implied by the pictorial representation. Levels 3 through 5 involve printed questions in a test booklet with the answers to be marked on a separate answer sheet.

Where Used: According to the most recent records, the tests of this series are being used at 24 Title VII Spanish bilingual program sites. These are:

Compton, California
Healdsburg, California
Olivehurst, California
Redwood City, California
Salinas, California
Denver, Colorado
Naples, Florida
Chicago, Illinois
Boston, Massachusetts
Springfield, Massachusetts
Albuquerque, New Mexico
Las Cruces, New Mexico
New York City, New York
Rochester, New York
Abernathy, Texas
Austin, Texas
Del Rio, Texas
Houston, Texas
La Joya, Texas
Laredo, Texas
McAllen, Texas
San Antonio, Texas
Zapata, Texas
Milwaukee, Wisconsin

Technical Data

Development of Spanish and English Editions of the Inter-American Series[1]

[1] The description to follow is a summary of the test author's claims about the development and intent of the Series. It does not reflect the reviewer's judgement about the test development, motivation for the Series, or content of the items. In a later section the reviewer will take issue with some of the claims made by the test author about the Series.

The uniqueness of this series of tests stems from the fact that there are so-called parallel forms, one in English and one in Spanish, yielding comparable scores. The entire Series of tests grew out of a study of teaching English in Puerto Rico in the 1940's. Both language editions of the tests were originally developed in Puerto Rico, after the notion of translating English tests into Spanish was rejected. According to the author(s) an attempt was made to develop parallel forms by bringing together native Spanish- and native English-speakers to construct the tests. The objective of this procedure was to select test items common to the two cultures and of similar difficulty. The tests were first published in 1950. The Inter-American Series to be examined here is the most recent version. Part of it was published in 1962 and part in 1966.

The goal in construction of the test pool items for both language editions was to create items with the following characteristics:

1. items common to, but not necessarily of the same frequency, the cultures of the Spanish-speaking and English-speaking peoples of the Western Hemisphere.
2. use of the same pictures, drawings and numbers in the non-language parts of the test booklets.
3. use of the same directions and same verbal content, expressed for one edition in standard English and for the other in standard Spanish of similar difficulty. The test developers claim that the Standard Spanish and English avoid local idioms as much as possible and that the tests have been designed for use without significant change wherever they may be administered.

Items from the item pool selected for inclusion in the test were chosen in the following way. Spanish items were administered to Spanish-speaking children and English items to English-speaking children. The relative diffi-

culty of each item was examined and its discrimination between the more and less able groups, as determined by total test scores, was noted. Only those items which discriminated between the more and less able groups in both language groups and which conformed to the previously mentioned specifications were selected for the published edition.

With the history of the development of the Series in mind, norms, reliability and validity and other technical aspects of the tests can be examined.

Norms

The test author and publisher take a unique position with respect to normative data on the Series. They recommend that the tests be used with regional or local norms (as contrasted with national norms) "to be prepared by those who use the tests." With respect to the original sample on whom the Series was developed, there is little information beyond the fact that it contained English and Spanish-speaking children, presumably in Puerto Rico.

What the author and publisher do provide are: (1) some norms, presented incidentally, based on data provided by some test users; (2) some estimations of norms based on calibration of the Series with other standardized tests with published norms (equivalent scores method) also provided by test users; (3) detailed instructions with respect to developing local norms and to calibrating the Series with other tests. Some of the tests which have been calibrated with the Series are:

Tests of General Ability, Level 1 with Goodenough-Harris Draw-A-Man

Tests of General Ability and Tests of Reading, Level 5, English, Form CE with some Project Talent tests

Tests of General Ability, Level 5 with one administration of the College Board Scholastic Aptitude Test at University of Texas

Estimation of "national" norms for Levels 3, 4, 5 of both ability and reading tests through calibration with various Educational Testing Service Tests.

Some of the Spanish edition of the Series with a test developed by the Puerto Rico Department of Education.

Percentile scores yielded by various levels of the test in some Spanish speaking countries (e.g. Mexico, Panama, Venezuela, Costa Rica, Chile) are also provided.

In examining the section of the manual which deals with the normative data, it becomes clear that there is a great deal of space devoted to normative performance. However, it is impossible to summarize or conclude anything from the data for they are in no way systematic. They are merely the performance of various groups, from various parts of the world, on various tests of the Inter-American Series. Furthermore, there are also no evaluations of the tests with which the Series are calibrated. From a practical point of view, little of the data provide a potential user with any helpful information. An individual test user should be prepared to construct his own norms.

Reliability

The indices of reliability provided in the manual are based on administrations of the two forms of the test (CE and DE) to the same groups of children after a "relatively short interval." Rather than reproduce the pages of tables illustrating the reliability coefficients of the various levels of the test, these coefficients will be summarized by noting the range of these coefficients.

Tests of General Ability
English Edition

Level 1	0.57 to 0.89
Level 2	0.53 to 0.82
Level 3	0.67 to 0.90
Level 4	0.41 to 0.82
Level 5	NOT PROVIDED

Tests of General Ability Spanish Edition	
Level 1	0.45 to 0.89
Level 2	0.49 to 0.80
Level 3	0.74 to 0.83
Level 4	0.74 to 0.88
Level 5	0.76 to 0.90
Tests of Reading Spanish Edition	
Level 1	0.79 to 0.86
Level 2	0.42 to 0.74
Level 3	0.64 to 0.90
Level 4	0.65 to 0.87
Level 5	0.48 to 0.82
Tests of Reading English Edition	
Level 1	0.84 to 0.95
Level 2	0.65 to 0.90
Level 3	0.78 to 0.95
Level 4	0.72 to 0.91
Level 5	0.74 to 0.93

Validity

There is no direct presentation or examination of the validity of any aspect of the Inter-American Series in the published manuals. One must infer the answer to the question of validity of the measures from material presented as correlations of the Series with other tests. Though not presented as material from which to infer validity, the correlational material

is massive (it comprises 22 pages in the technical manual). If it is not intended that validity be inferred from this material, then it must be said that the author and publishers of this Series of tests have presented absolutely no consideration of the validity of their instrument. A list of the tests with which certain levels of the Series have been correlated follows:

1.	Goodenough-Harris Draw-A-Man	(with English edition only)
2.	Metropolitan Readiness Test	(with English edition only)
3.	Otis Quick-Scoring Mental Ability Test, Alpha	(" " " ")
4.	School and College Ability Tests (SCAT)	(" " " ")
5.	Differential Aptitude Tests	(" " " ")
6.	Metropolitan Reading Test	(" " " ")
7.	Stanford Achievement Tests, Primary II, Reading	(" " " ")
8.	California Mental Maturity Test	(" " " ")
9.	California Achievement Tests Sequential Tests of Educational Progress (STEP)	(" " " ")
10.	STEP Reading	(" " " ")
11.	Project Talent tests	(" " " ")
12.	College Board Scholastic Aptitude Test (English and Spanish)	(with English and Spanish editions)
13.	Iowa Tests of Basic Skills	(with English edition only)
14.	Metropolitan Achievement Tests	(" " " ")
15.	SRA Achievement Tests	(" " " ")
16.	Iowa Test of Educational Development	(" " " ")

The tests of General Ability and Tests of Reading have also been correlated with each other and with teachers' marks.

The inadequacy and irresponsibility of this attempt at test validation cannot be over-emphasized. There is no attempt to generate a systematic defense of the validity of the Series or its psychometric structure. There is not even a reference made to the theoretical underpinnings of the Series, i.e. why certain questions were thought to be indicative of general or reading ability. The correlations with other tests (as listed above) do not fulfill any criteria for validity.

Furthermore, though the list of tests with which the Series has been correlated is lengthy it provides almost no information as it is not a systematic correlational procedure. The correlations are based on data from different levels of the Series collected on highly varied samples, located in different geographical areas. Sometimes the other test is correlated with several levels of the Series, sometimes with only one level of the Series. thus the data do not stand as a coherent entity. Finally, to infer validity of one test from its correlation with a second test implies the validity of the second test. And, as the test developers of the Series do not provide any information with respect to the validities of the other tests, the provided correlations are meaningless. That validity is not a requisite for published tests is a well-known fact. It lends additional support to the contention that the correlations of the Series with other standardized tests are a meaningless gesture.

Evaluations

A. The Center for the Study of Evaluation (UCLA), on a three-step continuum from good to poor, rated several levels of the Series for use with first, third, fifth, and sixth grades in the following way:

Tests of General Ability

	Grade 1		Grade 3		Grade 5			Grade 6		
	Verb. Num.	Non-Verb.	Non-Verb.	Total	Num.	Verb.	Non-Verb.	Num.	Verb.	Non-Verb.
Measurement Validity	poor	fair	fair	fair	fair	fair	fair	fair	fair	fair
Examinee Appropriateness	fair	fair	fair	fair	fair	fair	fair	fair	fair	fair
Administrative Usability	fair	good	good	fair	fair	fair	fair	fair	fair	fair
Normed Technical Excellence	poor	fair	fair	poor	fair	poor	fair	fair	poor	fair

Tests of Reading

	Grade 1	Grade 5	Grade 6
	Vocabulary	Vocabulary	Vocabulary
Measurement Validity	fair	fair	fair
Examinee Appropriateness	fair	fair	fair
Administrative Usability	fair	fair	fair
Normed Technical Excellence	poor	poor	poor

B. Summaries of reviews from The Fourth Mental Measurements Yearbook (Buros, 1953).

Tests of General Ability

Drake. Drake indicates general disapproval of the tests and advises that they only be used with extreme caution. He feels the only justification for the tests, in light of the many standardized tests of capacity, is its parallel Spanish and English forms. Even though the test is claimed to be culture-free, he questions this aspect of the test. Drake wonders if the tests are really culture-free and if both editions are equivalent in difficulty. If they are, then why, according to the provided median scores, is the ability

of children of the United States greater than the children of Mexico which in turn is greater than the capacity of the children of Puerto Rico? Drake questions these results and further asks if one can assume the sampling was equivalent in all three countries, if the motivation of the children was equivalent in all three countries, etcetera.

Durost. This review is not very helpful as it refers to information not currently available to the public, information privately obtained by Durost from the test author and/or publisher, information referred to in the review as graduate work being conducted at the University of Texas, or information which is only in older versions of the test manuals. Furthermore, several of Durost's comments are unclear. For example, in dealing with the validity of the test Durost says: "From the point of view of validity, it seems clear that these tests are superior to tests currently available in the United States for measuring mental ability for Spanish-English groups." One questions the validity data on which this statement is based and, further, what a "Spanish-English" group is.

With respect to validity, Durost's summary position is that the validity data, i.e., the correlations of the test with achievement tests, indicates that they fall within the typical range of such values, and that they provide no basis for thinking these tests are better than others. His criticism with respect to the tests norms is that they are of little practical use. Concerning mechanical details, Durost indicates that the art work of the test is not very good, that sometimes the intent of the picture is hard to determine, and that the separate answer sheet is awkward. In concluding his review, Durost says that there is nothing about the Tests of General Ability that would cause one to use them in place of widely-used standardized IQ measures. However, he indicates that its use with bilingual children at borders of English- and Spanish-

speaking countries is certainly desireable. He feels that the test represents the best that is available for use in Spanish-speaking countries. Durost ends his review with the hope that additional research will be conducted out on this test.

Tests of Reading

Orleans. Many of Orleans criticisms appear to be specific to the earlier version of the Tests of Reading. Nevertheless those remarks addressed to the validity of the tests and to the pictorial presentation appear to be still valid and will be discussed below.

First, Orleans is concerned with the validity of the tests and the context in which they are presented. He questions the context of both the English and Spanish editions. More importantly, he questions whether a context appropriate for measuring reading achievement "in English of American children" is also appropriate for measuring reading achievement in "Spanish of Spanish-speaking children." Orleans notes that there is no supporting evidence of content validity for either of both editions of the test, which further compounds the issue of validity.

As an example of the validity problem, Orleans cites an item where a picture of a woman washing clothes is followed in the English edition by the words wash, wake, walk, and call and in the Spanish edition by the words lavar, despertar, andar, and llamar. In the English edition, a child is required to distinguish between three words beginning with the same letter, all having the same number of letters in the word, while the same is not true for the Spanish translation. The effect of these circumstances on the validity (and reliability) of the test appears to have not been considered by the test authors and publishers.

As to the quality of the pictorial presentation of the reading tests, Orleans comments that they are poor and confusing.

Westover. (reviewed the English edition only). With respect to the English edition, Westover finds fault with the following aspects of the Tests of Reading. First, he finds the illustrations and format of a poor quality. Second, he remarks that the vocabulary section gives the test user little information regarding the pupils word-recognition skills. Third, he finds the tests do not provide enough information as they measure only two aspects of reading: vocabulary and comprehension (this latter problem appears to have been remedied in later editions of the test). He feels the tests require the addition of some measure of reading speed.

Irrespective of these faults, Westover feels the tests have face validity and that the materials are intrinsically interesting. He feels the tests' specific value are when used in connection with the Spanish edition in order to compare performance. Otherwise, he feels older and established tests of reading have more to offer the test user especially as they provide more adequate norms, data concerning reliability and diagnostic information.

Relavence of Tests for Spanish-Speaking Populations

There are several issues which must be considered in evaluating the appropriateness of the Inter-American Series of tests for Spanish-speaking populations. Of the more timely of these are the following: the value of the Series as an estimate of the ability or capacity of Spanish-speaking children; the value of the parallel forms with respect to culture-fairness or freedom from cultural bias; and the issue concerning the determination of which language edition is appropriate for usage with Spanish-speaking children of the United States.

The first of these issues concerns the accuracy of the estimate of ability provided by Series test scores. The question concerns the accuracy of the Series as a measuring device. Information contained in other sections

of this review (Technical Data) indicates that the Series has some very serious deficiencies. The investigation of its technical properties, i.e., the reliability and validity gives the impression of being confused, sporadic and random and does not impart the feeling that the Series is either reliable or valid. This feeling is borne out by ratings received by the Series from the CSE evaluators.

Leaving these technical matters aside, the Series has some major shortcomings on a much more basic level (which, or course, ultimately contribute to the Series' lack of reliability and validity). These concern the practical aspects of the test such as language and content, visual presentation and timing.

With respect to the language, careful examination of both of the language editions reveals the following problems. First, the directions are unclear and stilted and the word usage is awkward. The English directions contain such situations as the following. In the Test of General Ability Level 1 Association section, the directions state: "Now look at the hat in the next row. Put your finger on the hat. To which of the other pictures does the hat belong?" (Emphasis mine.) The Spanish directions, in what may be proper Puerto Rican Spanish are a poor choice of words from the point of view of Southwestern United States Spanish speakers. For example, the instructions refer to fila. Some bilingual educators point out that cuadro, linea or cerro would be better. Also, Level 2 Analogies (Test of General Ability) the instructions state "Estos dos dibujos son el primer par..." One bilingual teacher has suggested that the more appropriate Spanish phrase for the Southwest atleast, is "Estos dos dibjuos estan en pares."

The language of the stimulus materials is also of a troublesome nature. The words are a poor sampling of words in common usage and the choice appears

to be biased towards the words to which an upper-middle class child would have the greatest probability of being exposed. For example, Test of General Ability, Level 1, form CE, #17 "...find the warrior." "...busquen el guerrero." Also, Test of General Ability, Level 2, CE, #22, "...the picture which makes you think of refuge." "...del dibujo que les haga pensar enrefugio."

When examining the content of the test items (although not completely independent of the language in which they are expressed), one again finds the problem of situations which lack words of common usage in addition to ones of ambiguity (where more than one answer could be correct). For example, in Test of General Ability, Level 2 DE, grades 2-3 #3, the correct picture is the one of a fairy, "la hada." However, the concept of a fairy is not a culturally appropriate one. In item 15 of the same test, the stimulus word is unconscious, "inconsciente." The correct picture shows a man (sleeping?) on a couch. Item 11 of the same test asks to mark "...debajo del dibujo que les haga pensar en entrando solo," and shows two pictures of a boy alone at a door. In one picture the boy is knocking and in the other he is actually crossing the threshold.

Concerning the visual presentation which is identical for both language editions, there is criticism from many sources with respect to the poor format. The illustrations are crowded and small. Sometimes finding the right answer is dependent on finding a smile on a figure which is in one of 48 drawings on a 8 1/2 x 11 inch page, the smile being smaller than 1/32nd of an inch. Furthermore, the illustrations are line sketches, leaving much to inference and imagination. The spacing is very poor. Often in a series of drawings for an item, each drawing involves more than one person. In these cases it is difficult to tell which drawing the many people are supposed to be a part of. Finally the pictures are ambiguous, making it difficult to discriminate between chicks and birds, cups and glasses, a book and a box of kleenex, etcetera.

The next issue having relavence for Spanish-speaking populations is that of parallel forms. This is a key issue since the developers of the Inter-American Series claim they have parallel forms: an English edition and a Spanish edition. Examination of both versions, however, makes it clear that the Spanish version is a straight-forward literal translation of the English version and not a parallel form. (That it is not an English translation of the Spanish version is apparent from the nature of the illustrations and the cultural content of the items.) Parallel tests, technically speaking, measure the same psychological entity but utilize different sets of operations (i.e., items or tasks). A literal translation from one language to another does not fulfill this criteria. Given that the parallel form notion of the Inter-American Series is rejected, the cultural appropriateness of the Spanish edition becomes a major concern.

The cultural appropriateness of the Series is a serious issue because it superficially appears to be appropriate as it is in Spanish and is claimed to be a parallel form (and not just a translation). Such claims lend to a more ready acceptance of it by educators than of other tests with or without a Spanish translation. Thus the Series is potentially dangerous in that educators often assume they have chosen a valid test, given the Spanish parallel form, and will investigate the test no further. Unfortunately, the test does not even have much merit as a Spanish test, when one takes into account its upper-middle class Anglo-Saxon bias and its use of Puerto Rican Spanish.

The consideration of cultural appropriateness gives rise to the question of which children should get which edition of the test. Should Spanish-speaking children get the English or Spanish version? Which version should Spanish-surnamed children get? The crux of the issue is that Spanish-surnamed children cannot necessarily understand, speak, or read Spanish. Children

who can speak Spanish and understand spoken Spanish cannot necessarily read it. While these are obvious truths, they are not universally known. Some school districts give the Spanish version of the Series to all Spanish-surnamed children and think they are being very tolerant and culturally democratic in doing so. Other schools give the Spanish version to all Spanish-speaking children, again with the conviction that they are being sensitive to the needs of these children and are giving them the maximum opportunity to perform well. But, to pass Levels 1 and 2, it is necessary to understand spoken Spanish. To pass Levels 3 through 5, one much be able to read Spanish. And, of course, all of the reading tests require the reading of Spanish. Just how many Spanish-speaking children of the Southwest can read Spanish well enough to pass tests designed to assess an illusive an entity as their intellectual capacity?

The other side of the question is the appropriateness for children of a Spanish-speaking culture of the English version, with its stilted language, with its old-fashioned, Eastern U.S. wearing apparel, with its Anglo-Saxon characters, with its ambiguous questions, and with its poor illustrations. One must conclude that the appropriateness and value of the Series, in any language, for any group is questionable.

Reviewer's Remarks

In examining the Series one must be concerned about its reliability and validity. A problem in examining a series of tests as large as the Inter-American Series is that many poor technical properties tend to be overlooked due to the sheer bulk of the information presented (subtests, totals, forms, levels, etc.). Thus, although a quick appraisal of the reliability of the Series reveals that there appears to be much data on it, a closer examination reveals that it is impossible to summarize the reliability data as it is unsystematic and that much work is lacking on the reliability of all parts (and forms and levels) of the Series. Whereas a reliability coefficient of 0.45 does not make much impact when on a page of 50 coefficients, it does make a severe impact on the life of a child who has to take such a (sub) test. Such a coefficient is unacceptable as it indicates the test is unstable and inconsistent.

With respect to the validity the reader is referred to the remarks on page 9. In summary, it may be said that the test author(s) and publishers have been grossly negligent in making available a test on such a large scale which has no validation. One wonders what the test constructors thought they were doing.

The problems arising from the lack of investigation of reliability and validity are greatly magnified by the existence of the alleged Spanish parallel form. This form, in combination with its availability in levels covering preschool to grade 13 makes the Series highly attractive to educators. However, in light of the fact that the Spanish form is not parallel, that the Spanish language usage is poor, and that the reliability and validity are so lacking, one can see what a deceptive test the Inter-American Series actually is.

PART III

ABSTRACT

of

"A SYSTEM FOR CRITERION-REFERENCED ASSESSMENT OF A BILINGUAL CURRICULUM" by Eduardo A. Apodaca director Project Hacer Vida, Title VII Bilingual Education

This pioneering effort by the staff of Title VII Bilingual Education Project "Hacer Vida" began in April of 1971. At the time, no testing alternatives existed for the project that was dissatisfied with standardized tests. The choice was always one of which standardized measure would be used. The overwhelming majority of these instruments are designed to measure competencies in the English language.

Another inconsistency in the initial evaluation method used resulted in trying to measure the achievement of performance objectives through the use of standardized instruments. There was a lack of correlation between what the tests were testing, and what the teachers were actually teaching. It came as no surprise to anyone when the six participating superintendents voted to eliminate all standardized tests from the 1971-72 Evaluation Design. The Berkeley Conference presentation is in effect a 'blow-by-blow' description of the events that have been experienced by project personnel in designing an evaluation alternative to norm-referenced measures.

The "Hacer Vida" Criterion-Referenced Model has implications for other bilingual education projects but also to traditional programs. The project is in the process of implementing an instructional system based on performance objectives for both the Spanish and English curriculum. Project staff are participating in the design of instruments that actually can test what is being taught. English Criterion-Referenced Instruments have been developed for first and second grades in the areas of Language Arts and Math. A Spanish Criterion-Referenced Instrument has also been created for use in both first and second grades. Teachers involved in this effort have been continuously refining their product.

One of the most valuable "spin-off" benefits has been the participation by teachers in determining what accountability model they will have to teach by. Teachers in the program have; in effect,designed the tests they are being evaluated by.

A unique feature of this criterion-referenced assessment model is the utilization of a student assessment card based on the McBee Keysort System. As students accomplish objectives, their card is punched. A group of 30 cards can be easily sorted with a needle to pull out groups of students that have not met the desired objective.

Another component of the evaluation model is the performance objective box. First grade teachers worked as a team in compiling their own box of objectives. Second grade teachers worked in similar fashion to organize theirs. The steps that were taken in coming to agreement on a set of objectives were: A. Research and review of all available performance objective models, such as the IOX Objective Bank. B. Selection of objective clusters; C. Concurrence on final selection; D. Revision of selected objectives onto the "Hacer Vida" Objective Card Format. E. Identification of optional procedures that could be used in teaching each objective; F. Citing textbook references on each objective card to link each objective with appropriate lessons.

A publication entitled "A System for Criterion-Referenced Assessment of a Bilingual Curriculum" by Eduardo A. Apodaca is currently available at a nominal fee from Title VII Project Hacer Vida, Office of Riverside County Superintendent of Schools 46-209 Oasis St. Indio, Calif. 92201. Statistical information on the criterion-referenced instruments will be available on the 1971-72 Final Evaluation Report, to be published by August 1972.

Eduardo A. Apodaca, Director
Project Hacer Vida, Title VII-ESEA
Office of Riverside County Superintendent of Schools
46-209 Oasis St.
Indio, Calif. 92201
(714) 347-8511 ext. 313

PART IV

SOME CAUTIONARY NOTES ON ATTEMPTING TO ADAPT IQ TESTS FOR USE WITH MINORITY CHILDREN AND A NEOPIAGETIAN APPROACH TO INTELLECTUAL ASSESSMENT: PARTIAL REPORT OF PRELIMINARY FINDINGS*

Edward A. De Avila

Multilingual Assessment Program
(Joe R. Ulibarri - Project Director)

Traditional tests of intelligence are inappropriate for the minority child. They are particularly inappropriate for those who come from non-English speaking backgrounds. Such diverse groups as the popular press, the courts, civil rights organizations as well as state and federal agencies have all been involved in pointing to the failure of the testing industry to fully consider the cultural and linguistic differences of minority children when constructing, publishing and selling these tests.

Since the industry stands to gain increased revenues through the use of its materials in federally-supported programs, it has responded to this criticism by:

1) translating existing intelligence tests for non-English speaking children

2) adjusting norms for ethnic sub-groups

3) attempting to construct culture-free tests

There are distinct problems with each of these approaches.

*Acknowledgement

The project directors would like to acknowledge their special indebtedness to George McCormick, Principal of Hazelton School in Stockton, California. Special thanks are also given to Gregorio Rios and Toni Castillo, who assisted in the data collection, to Marvin Hanely for his special help in analyzing the data, the entire staff of the Multilingual Assessment Program, and to Stanley France who assisted in the preparation of the manuscript. Finally, we would like to extend our deep appreciation to Juan Pascual-Leone of York University in Toronto, Canada for the use of his Figural Intersection and Water Level tests.

With respect to translations, several problems arise. First, regional differences within a language make it difficult either to use a single translation or to compare across different translations. Thus while the word "tostone" refers to a quarter or a half a dollar for a Chicano child, for a Puerto Rican it refers to a squashed section of bananna which has been fried. Second, the assumption that non-English speaking children speak one language exclusively lends to mono-lingual translations which, in many cases, are not related to the actual spoken language of the child which may be a combination of languages. This assumption leads to the further assumption that, because a given language is the spoken language it is also the "written language." One finds many examples of tests written in Spanish being given to Chicano children who may speak Spanish but who have had absolutely no prior instruction in reading Spanish. Third, another problem in translating tests is that words in one language have frequencies and potencies which generally cannot be compensated for in a direct translation to a second language. In other words, having a cognate is no guarantee that it is used in the second language with the same frequency as it is used in the first language. For example, the word "pet" is a common word in English yet, its Spanish cognate, "animal domestico," is almost never used. A related problem in this context has to do with the fact that translating a word from one language to another can vastly alter its meaning. Thus, there are wide varieties of seemingly harmless English words which translate into Spanish swear words or "palabras verdes." This being the case, translating a large egg into a "huevon" may satisfy grammatical requirements and seem harmless to an Anglo

translator, it nevertheless fails to consider that portion of the word's meaning which "does not translate." Fourth, straight forward translations of existing tests represent a complete denial of cultural differences. In many cases this leads not only to unfair tests but to tests which require the child to break from his own cultural tradition. Thus, asking an Indian child "who discovered America" or asking a Moslem child to "draw a man" requires not only that the child break with cultural and religious tradition but also that he set himself apart from his own reference group.

The second major response of the testing industry to criticism with respect to the testing of minority children has been to establish regional and ethnic norms; in other words, simply to lower the criterion levels on the basis of ethnicity. This leads to expecting less from the brown, black or lower socio-economic white students than from the middle class Anglo child. Awarding "bonus points" to minority children to compensate them for their "deprived background" is based on the same proposition as lowering norms. It is nevertheless a simple-minded solution to gratuitously award Chicano children extra points "because they speak a little Spanish."

These practices are all based on the common notion that ethnic norms should be established. Such practices are potentially dangerous because they would provide a basis for invidiously determined comparisons between different racial groups. The tendency would then be to assume that lower scores are ultimately indicative of lower potential and would not only continue the self-fulfilling prophecy of lower expectation for minorities but would also reinforce the genetic inferiority argument advanced by Jensen (1958), Shockley (1971) and others.

Third, there is a problem which cuts across these issues which in many cases may negate attempts to "clean up the tests." This problem involves validating a test of intelligence by correlating it with measures of achievement. The assumption is that the brighter the child, the greater his achievement. This appears reasonable enough, for certainly if a child has a high capacity, it must be related to some sort of achievement. With respect to the minority child, however, the relation between intelligence and achievement breaks down. It is a notorious fact that traditional curriculum has little relevance to the minority child. As such, any attempt to validate intelligence tests for these children by relating them to traditional curriculum is doomed to failure because a bright Chicano or Black child does not necessarily thrive on a curriculum designed for a mid-western Anglo population.

The fourth major difficulty in the testing of non-Anglo children is the false assumption that a test can be constructed which is independent of culture. Such a test is difficult if not impossible to construct. Consider that a culture must inevitably be defined by a particular set of referents. Intellectual activity must per force refer to the manipulation of these referents. As such, intellectual activity or any mental operation must involve the processing of information, that is, referents defining an environment or culture, which, by definition defines that particular culture. Aside from the problems inherent in depicting a culture without a referent, to ignore this problem would be to recapitulate the problems in Descartes's assumption that objectless (without a referent) is possible.

Over and beyond these problems, an analysis of the content and format

of items used in a large number of traditional IQ tests reveals several highly interrelated types of items suggesting that the tests are measuring something other than that for which they were designed. Traditional IQ measures may therefore also be described as measures of socialization, productivity or level of aspiration, specific experience and endurance. Consider the following as only a few of the possible illustrations that can be mentioned.

Socialization. Items of this type draw primarily on the nature of one's socialization and are couched in such a way as to actually be measures of the child's family value system. The referent system, is of course, the dominant Anglo middle class. The confounding effects of this problem are particularly evident in the "comprehension" scale of the Weschler (WISC) where children are asked such questions as:

> "What is the thing to do if you lose one of your friend's toys?" or
>
> "What is the thing to do if a fellow much smaller than yourself starts a fight?"

Allowing for the stilted manner in which the question is phrased and assuming that the child knows all of the vocabulary, it still seems perfectly obvious that this type of question has little or nothing to do with a child's ability to process, manipulate or code information but, rather with whether he has been socialized under the particular ethical system implied by the question.

Productivity or level of aspiration. Many tests confound what they hope to measure with a measure of productivity or level of aspiration. For example, in a large number of tests the child who produces the largest number

of responses is rewarded whereas, the child who (for whatever reason,) produces fewer, is punished by receiving a lower score. Thus, in the Draw-A-Man, the child who produces the more elaborate figure receives the higher score. The problem here stems from an assumption that all subjects will produce as many responses as they are able, i.e. have the same level of aspiration. The effects of this assumption are particularly evident in timed tests, which constitute the majority of published tests. In these tests children are required to "work quickly and efficiently" without regard for the child who is simply not in a hurry nor particularly motivated to be so.

Another type of test which may be grouped under this category is the "endurance test." This particular type of test, for purposes of boosting statistical reliability, requires that the child answer a large number of questions which vary little in content. This problem is particularly evident in the group tests such as the Lorge-Thorndike Intelligence Test and the California Test Bureau Series.

Experience of specific learning. In tests which require subjects to answer questions of fact, there is an implicit assumption that the children taking the test will have had a more or less even chance of having been exposed to the fact being tested by the question. The spuriousness of this assumption is witnessed by any number of examples where children are asked questions of vocabulary. Granted a high positive correlation between intelligence and vocabulary, it is impossible, nevertheless, to determine whether a minority child has missed a test item because he lacks the capacity to understand a given word or because he

simply has never been exposed to the word, e.g., "nitroglycerine" (in the WISC), "fire hydrant" (in the Betty Caldwell and Peabody) or "crevice" (in the Otis-Lennon).

The fundamental problem with most of the tests mentioned above and, indeed IQ tests in general, is that test publishers have failed to fully consider the problems associated with testing the minority child. Moreover, it would seem that the attempts to deal with these problems by the above mentioned means will lead to limited success for the reasons discussed. However, since the results of tests are used to determine the educational and, by extension economic and social future of school-age children it, therefore, behooves test publishers to more fully consider the minority child's cultural background. A publisher who has considered cultural background would know, for example, that the Chicano child is reluctant to guess when he doesn't know the answer to a question; that the Indian child is taught in the spirit of cooperation rather than competition and is reluctant to compete with his peers; that Black, Chicano and Indian children have little experience in developing test-taking strategies which would enhance their performance; and finally, that there are a significant number of children from all of these groups who view the schools as threatening, hostile and alien.

In summary, it may be said that the major problem in the psychometric approach to intelligence testing described in the previous notes is that environmental factors such as linguistic and cultural differences have not been taken into account. The position to be taken here, in contrast to the psychometric approach, would argue, in agreement with Piaget

that the determination of intelligence must be studied through the examination of intra-individual rather than inter-individual approaches. Thus in the present view, intellectual development is characterized by the extent of internal control of functioning versus external control of functioning at any given stage of development.

With the understanding that testing procedures must distinguish between external-environmental and internal-developmental variables, the determination of a subject's intellectual development thus becomes a two-step process. In the first step it becomes necessary to remove the effects of these external factors before actually testing the subject. The second step involves a determination of the extent of internal variables through the use of tasks which vary in the degree of control required to produce a correct response.

The use of a "experimental repertorie control (ERC)" provides for the control of external variables which can reflect diverse experential and stylistic differences rather than differences in intellectual capacity or internal control of functioning. The application of a controlled repertoire in which subject differences are removed through pretraining procedures has been attempted by Pascual-Leone & Smith (1969), Pascual-Leone (1970) and De Avila (1971).

Pascual-Leone (1970) used a variety of the Piagetian tasks and the Witkin et. al. (1962) measures of field dependence-field independence in a factor analytic study of cognitive development and cognitive style. An essential feature to Pascual-Leone's procedures is that prior learning is used as a control variable rather than as a dependent variable (see Pascual-

Leone & Smith, 1969). Using prior learning as a control, Pascual-Leone (1969) found highly stable results across a number of Piagetian tasks. In discussing the failure of previous experimenters to obtain high correlations among Piaget's tasks, Pascual-Leone (1970) notes that these poor results may be due to (1) poor reliabilities caused by the small number of items per test, (2) failure in "relevant linguistic pretraining," and (3) failure to note that subjects do not always function at their "structural" or highest level of operativity.

In another study by De Avila (1971) using upper-middle class children, it was found that when the background of the subjects was controlled through the use of experimental control tasks, low correlations were found between a standardized intelligence test, (the Otis-Lennon) and a number of Piagetian tasks. Such results imply that the IQ measure may be highly related to the external variables such as educational and social background. Moreover, these findings suggest that when these factors are controlled for through pretraining, IQ ceases to be an adequate measure of intellectual development. Replication of this finding with low socioeconomic subjects would support this position. More important to the current research is the purpose of establishing the reliability and construct validity of the current measures with a new respect to the Piagetian developmental hypothesis.

A second major purpose of the present study which replicates and expands the largely unpublished extensive results of Pascual-Leone, Parkinson and De Avila at York University and/or Boulder, Colorado was to examine the psychometric properties of several Piagetian tasks which vary according to the extent to which external variables are controlled. The third purpose

to which this research is directed is the issue of group administration of Piagetian tasks. Educational situations usually require group testing because of the large number of subjects involved relative to the manpower available. Piagetian tasks have historically been individually administered. However, Dodwell (1961) and Harker (1960) have shown that the child's conception of number can be tested in a group setting; De Avila, et. al. (1969, 1968) have measured several conservation tasks and spatial perspective problems in group situations. De Avila et. al. (1969) found adequate reliabilities for the conservation of substance and egocentricity measures, suggesting the further possibility of using Piagetian-based group measures to evaluate the developmental-psychometric properties of tests which are applicable across a broad range of development. Similarly, Pascual-Leone (1969 and Pascual-Leone & Parkinson unpublished) have adapted a number of Piagetian and neo-Piagetian tasks to group settings with a high degree of success.

The goals of the present research were thus:

1. To examine some of the relationships between the neo-Piagetian approach to developmental scaling and traditional approaches embodied in psychometric testing.
2. To test the applicability of the "experimental repertoire control" (ERC) concept as a procedure for testing minority children.
3. To test the feasibility of using Piagetian measures to determine the developmental levels of minority children.
4. To examine the psychometric properties of the Draw-A-Man and Columbia Mental Maturity Scale for minority children.

5. To examine the relationship between developmental and I.Q. analysis procedures for minority children.

Instruments

Four Piagetian or neo-Piagetian tests were given: the Cartoon Conservation Scales,(De Avila, 1968a; 1968b; 1969), the Conservation of the horizontality of water as measured through the Water Level Task, (Pascual-Leone, 1966; 1970; Pascual-Leone & Parkinson, unpublished), the Figural Intersection Task (Pascual-Leone, unpublished; Pascual-Leone & Smith, 1969), and the Serial Task (De Avila, 1971). In addition two standard measures of intelligence, the Columbia Mental Maturity Scale and the Draw-A-Man were also used. Each of these measures are briefly described below.

CARTOON CONSERVATION SCALES (CCS)

Several measures of Piaget's conservation tasks were assessed by means of the cartoon format developed by De Avila et. al. (1968a; 1968b; 1969). In De Avila's procedure, three cartoon frames are presented in which two children discuss a Piagetian task. In the first frame an equality is established between two objects according to the dimension being studied (i.e., number, length, substance, etc.). In the second frame an identity transformation is depicted and in the third frame the question of conservation of equivalence is asked. On the right side of the panel three possible answers are presented. The three alternatives which show the characters responding to the question are randomly ordered as to correctness in order to avoid position effects. Similarly, wording is altered from item to item in order to avoid the possible effects of acquiescence. Background on the conservation scales and an illustration of the dialogue from each scale are

presented below.

In its current form the CCS consisted of thirty cartoon panels. There were six examples of five tasks. The panels were presented to the subjects and the story line was read and elaborated upon in order to facilitate understanding of the question. The subjects task was simply to mark the one (alternative) "that makes the story true."

Conservation of number is measured by showing blocks on a table. The dialogue is as follows: Frame One: "How many blocks are there?" Frame Two: "There are seven in each row. I'll put these in a bunch." Frame Three: "Are there fewer in the row than in the bunch?" There are three possible responses from which the child chooses his answer. Each alternative provides the child with written (i.e., child points to one, another, or to both sets of blocks.) As in all cases the child simply picks his answer by putting an "X" on the picture "that makes the story true." (See example 1)

Conservation of substance is measured through items such as the cartoon where the following dialogue takes place. Frame One: "These two clay balls are the same size." "They both have the same amount of clay." Frame Two: "I'll roll one into a long hot dog shape." Frame Three: "Does one have more clay than the other one now?" In the response frames the responses are: (boy points to both) "They have the same amount", (boy points to hot dog) "The hot dog has more", (boy points to ball) "The ball has more." (See example 2)

Conservation of surface performance on the task requires that a subject recognize that no matter where a given number of objects are located

on a surface, the amount of surface exposed remains the same. An illustration from the CCS uses a toy farm placed on a table. The dialogue in Frame One is: "See the little farm." "The cows are all over the table." In Frame Two the dialogue is: "The cows need to have more grass." "Put the buildings on the back of the table." In Frame Three the question is: "Is there more space on the table now?" The response order is: "There is less space now." "There is the same space." "There is more space now."

Conservation of weight in the CCS one of the illustrations involves two children balancing on a seesaw. In the first frame, the two children are shown from a distance and one says "Hey, this is fun. We can go up and down." In the next frame the second child says "Let's see what happens when we stop." In the third frame, the two children are shown in a balanced-horizontal position and one child asks, "What will happen if I lie down?" The three alternatives show the seesaw in several positions with the child who asked the question in a lying down position. It should be noted that the position of the child who is lying down is depicted in such a way as to indicate no change in the distance between himself and the fulcrum (seesaw center post) so as not to alter the leverage relationships. (See example 3)

Egocentricity In this measure, the subject is asked to picture how a setting would look from a perspective other than the one from which he is looking. One illustration from the CCS uses the concept of taking a picture of a toy barn, silo, and tractor as follows: "See my new camera." "Take a picture of my farm." "I'll take the picture from over here", (view opposite that of person who "owns" farm). Frame Three: "What will the picture look like?" The response frames show the picture taker's viewpoint, the "owner's"

viewpoint and a side view, each with the caption, "It will look like this." (See example 4)

WATER LEVEL TASK (WLT)

The conservation of the horizontality of water measure utilized here was introduced by Pascual-Leone (1966, 1970) as a standardized quantifiable version of the Piagetian test (Piaget & Inhelder, 1968). A more complete description of the relative parameters of this type of task can be found in the semantic-pragmatic analysis of the relative strengths of objects in the field done by Pascual-Leone (1970).

In this study, a special version of Pascual-Leone's group tests by Pascual-Leone & De Avila (1972) was used. Subjects were presented with individual booklets which contained five horizontal or vertical two-dimensional bottles, eight two-dimensional-tilted bottles and four three-dimensional bottles, two of which were also tilted. The subject was asked to draw a line where the top of the water would be if the bottle were half full and then to place an "X" in the part that contained the water.

FIGURAL INTERSECTIONS TEST (FIT)

The figural intersection test is a group administered paper-and-pencil test in which subjects are required to place a dot in the intersecting space of a varying number of geometrical figures. It was developed by Pascual-Leone and constitutes a figural analogue of Piaget's "Intersection of Classes" (1932). The type of overlapping figures utilized in this test were originally devised by Abelson (1911) for another purpose. In a series of unpublished studies, Pascual-Leone has shown the test to have a high degree of internal

consistency (split-half reliability = .89) as well as being significantly related to tests of similar logical structure (Pascual-Leone & Smith, 1969). For example, it has shown a high correlation with the WLT described above. Combined with the WLT, in the present context, it was taken as an index of developmental level. This relationship has been previously found in a series of unpublished studies by Pascual-Leone & Parkinson (1969).

SERIAL TASK (ST)

The serial task (De Avila, 1971) is a short term memory task which is individually administered in two phases. First, subjects are pre-exposed to the stimulus materials used in a second testing phase. In the pre-exposure or pre-training phase, each subject is shown a series of 10 different 35 mm. color slide transparencies of pictures depicting a donkey, house, airplane, etc. Subjects sit facing a screen situated on a wall six feet away. The 10 illustrations are presented by means of a Kodak 650 carrousel slide projector. To introduce the task, each subject is shown each figure and asked to give its name and color (i.e., "a yellow hat"). Following this initial introductory phase and after the subject was able to correctly identify each figure ten times when presented in rapid random succession, the testing phase was begun.

The test phase was conducted in a "free recall" manner (Adams, 1967) where, without any prior knowledge of the length of a list, the subject was asked to reproduce the list ignoring the order in which the individual items are presented. Subjects were shown a series of individually presented figures terminated by a blank slide, and asked to tell the experimenter what they saw. The exposure time for each individual slide was .750 msec.

There was no requirement that the sequence of the presentation be maintained, or that the subject respond within a specified period of time, or produce a predetermined number of responses. The child was simply asked to reproduce what he saw using whatever labels were convenient.

There were seven sets of figures presented to each subject. These seven sets varied as to the number of stimuli within a series. There were 28 sets in all, 4 consisting of one figure, 4 consisting of two figures, 4 consisting of three figures to 4 consisting of seven figures. The number of figures presented within a series, as well as the individual figures, were randomly varied. Finally each illustration was presented no more than once in a series.

DRAW-A-MAN & COLUMBIA MENTAL MATURITY SCALE

In addition to the CCS, FIT, WLT and ST, two standard measures of intelligence, the Draw-A-Man (D-A-M) and the Columbia Mental Maturity Scale (CMMS, Burgemeister, 1954) were included in the test battery. These measures served to establish some indication of the relationship between measures of intelligence currently in use with minority children and the above described measures.

Procedure

The CCS, FIT, and WLT, testing was conducted in small groups. For the D-A-M, CMMS and ST testing was done individually by one of two bilingual-bicultural experimenters. Where necessary, instructions and testing were carried out in Spanish.

Subjects

Subjects for the experiment were 100 first through sixth graders at a

central city school in a city of approximately 115,000 on the West Coast. Ethnic composition of the group was: 63.2% Mexican-American, 1.8% Black, 22.6% Caucasian, and 12.2% other non-white. Testing was done during two consecutive months.

Results

In order to establish the construct validity of the conservation measures included in the CCS, a principal components factor analysis with rotation was performed. Table 1 contains the factor loadings. A total of 50 percent of the variance in the matrix was accounted for by the solution. With the exception of two of the conservation of number items, the factors clearly represent the conservation measures included.

Refer to Table 1

Given the distinctness of the tasks, scale scores for each measure of conservation were obtained by simple summation. Table 2 shows the values of Cronbach's Alpha and the Kuder Richardson Formula 20's (KR-20) and homogeneity ratios (HR, see Scott, 1960) for each of these scales.

Refer to Table 2

Since the Water Level Task contained three different situations involving the conservation of the horizontality of water, it, too, was factor-analyzed. The results of this factor analysis are shown in Table 3. The first three factors had eigenvalues greater than one and accounted for a total of 65% of the variance in the matrix. Reliability data is shown in Table 4 for the

subscales and the total scale obtained by a simple summative basis.

Refer to Table 3 & 4

The reliabilities for the ST are shown in Table 5. The low reliabilities at the extreme low end of the scale are clearly due to the lack of variation of performance among subjects, as all subjects remembered the single picture.

Refer to Table 5

The FIT yielded reliabilities similar to those of the ST as may be seen in Table 6. However, little variation was found for the sets involving seven or eight figures.

Refer to Table 6

As is evident from Table 6, all the scales of the FIT were highly reliable with the exception of those at the extreme ends. Also, almost identical results were found using different measures of reliability.

High reliabilities were also found for the CMMS (Cronbach Alpha = .869 KR20 = .887), and for the D-A-M (Cronbach Alpha = .846). The homogeneity ratio for the CMMS was .124 and .116 for the D-A-M. According to Scott (1960), the homogeneity ratio is a conservative index of the average correlation between test items. In practice ratios between .150 and .600 are acceptable (personal communication with William A. Scott, University of Colorado, 1967). The lower the ratio, the more complex and heterogeneous is the concept. Values below .150 suggest each item is a measure of a different concept with the

test scale not measuring a unitary trait or concept. Thus, the homogeneity or internal consistency of both the D-A-M and CMMS would appear to be on the low end of the acceptable range. On the other hand, the somewhat depressed homogeneity of the ST (HR = .112) and FIT (HR = .199) would have been due to the low performance variability found at the extremes of the scales.

The intercorrelations of measures used in the study are shown in Table 7. Of particular interest are the negative relationships found between age and IQ and the lack of relationship between the two IQ measures.

Refer to Table 7

Further evidence of the inappropriateness of the psychometric IQ model as embodied in the CMMS was found in a factor analysis items. Of the first 50 items on the test, only a few items were missed yielding a sample mean of 49.84 with a standard deviation of 0.48 for these items. From the lack of discrimination among subjects it appears that these items are worthless in this situation. The last 50 items were factor-analyzed by the principal components method and varimax rotated. The first factor accounted for 15% of the variance while 14 factors had eigenvalues greater than one. A varimax rotation was performed on the first five factors. Of the 50 items, 31 had loadings of .400 or better, 12 on the first factor, 5 on both the second and third factor, 4 on the fourth and 6 on the fifth factor. An examination of the items suggests no consistency of conceptual operation for a given factor. Factor one, for example, contains functional analogies, class exclusion, number analogies, and size analogy items. The low communalities and the fact that the five factors accounted for only 34 percent of the total variance further

suggests difficulty in interpretation of the instrument.

Refer to Table 8

In the otype procedure described by Tryon and Bailey (1971) an attempt is made to identify groups of subjects which have similar test profiles. In a process somewhat like template matching used in pattern perception studies (Uhr 1963) subjects are grouped according to response patterns. According to the hypothesis that different age groups will pass different conservation tasks one would therefore expect to find different groups of subjects to have similarly responded to the different conservation tasks.

Since the different Conservation tasks measured by the CCS are assumed to be mastered at different ages, it was hypothesized that children could be grouped according to otype performance differences which would be reflected by statistically significant differences between them. Thus conservation of number and surface should be mastered by all subjects, substance by all but the youngest group and ego and weight only by the oldest subjects. In order to test this hypothesis, the procedure described by Tryon and Bailey (1970) as the otype approach was used with a modification to allow for testing the structure of the types against hypothesized types. In the procedure used, T-scores were computed which reflected the number of items per conservation scale which a subject would have to attain to be reasonably sure of being able to perform the task (in this case 4 out of 6 were used) and also the number of scores (2 out of 6) for a chance response were computed (2 out of 6) as a T score. The expected types were established to reflect the order in which the concepts were supposed to be attained. In

the case of the WLT, ST, and FIT, 10 T-score points were arbitrarily used between each of the types. This procedure resulted in three hypothesized types, which are shown as the first entry in Table 9. In the computational procedure, all scores are first converted to T-scores. The distance of

---

Refer to Table 9

---

each subject's scores from the means of each of the arbitrary types is then computed using the least-squares approach, and divided by the number of variables. The subject is assigned to the type from which the distance is the smallest providing that distance is not greater than a predetermined control (in this case 11 T-score points). Once this is accomplished for each subject, new means are generated and reiteration begins with recomputing distances. This continues until there are no changes in type membership.

The types which resulted appear to be quite homogeneous. Further, it should be noted that there were no reversals of what in trend with the order of increment being from otype 1 to otype 3. Mean t-tests of differences from the expected types are shown as the last entry in the table. Those tasks (WLT, FIT, ST) less subject to environmental influence seemed to match the arbitrary types more closely than the conservation tasks.

Analysis of variance and independent t-tests were computed for the final otypes for all variables included as well as for age. The results of this analysis are shown in Table 10. Of the more environmentally-independent

---

Refer to Table 10

---

measures, only one comparison failed to find a difference. This failure to find a difference occurred between otype 2 and otype 3 on the ST. In attempting to determine why this resulted, it was found that one of the experimenters had repeated the stimulus to some subjects when so asked. This practice could have resulted in higher scores for the younger subjects of otype 2 due to practice effects.

Discussion

In general terms, the findings described above give support to the procedures utilized in the present approach. A possible criticism, however, stems from the limited sample size and a caution must therefore be taken in generalizing these findings to other larger populations. Similarly, these results are limited by the fact that the subjects represent a rather limited sampling of the urban-rural continuum. With these general limitations in mind the following will consist of a discussion of some of the more pertinent findings.

The first and perhaps most immediate conclusion to be drawn from the present research concerns the nature, structure and possible inappropriateness of the psychometric IQ model as embodied in the DAM and CMMS. While high reliabilities were found for both of these tests, the low homogeneity ratios indicate that both tests are tapping a somewhat more amorphous concept than general intelligence. Second, since both the DAM and CMMS showed negative correlations with age, one would have to consider that at least the age norms, if not the entire tests, are inappropriate for the present sample. Third, the fact that the correlation between the two tests was negligible similarly calls into question the procedures of these two tests. Finally, since the

factor analysis of the CMMS showed low overall communalities for the items and did not produce a factor structure consistent with the structure described in the test manual, there is little support for the test. In fact, the basic conclusion which must be drawn from these findings is that both the CMMS and DAM should be used with great discretion.

In contrast to these results, the CCS, WLT, FIT and ST results were more encouraging. The conservation scales (CCS) showed a high degree of internal consistency as indicated by the factor analysis structure as well as by the homogeneity and reliability indices. With the exception of the "surface" items, further support was provided for the CCS by the hi[illegible] correlation of the subscales with the other tasks. An examination of the age trends for the surface subscale showed it to have the lowest overall correlation with age ($\underline{r} = 0.212$). Since it was expected that all of the subjects would be able to pass items of this type, the overall correlation was expected to be low due to restricted variance (i.e., all subjects were correct). A similar finding was anticipated and found for the number subscale. However, the mean probability of a correct response for the number subscale items was .84 whereas, it was .26 for the surface subscale. This finding is in sharp contrast with the anticipated result and raises questions as to the applicability of the cartoon format with this type of item as well as with Piaget's analy[illegible] of the task. Certainly, since the mean probability of a correct response was below chance (.33) there was a great tendency on the part of all subjects to "centrate" on misleading cue provided in the item. This findings is consistent with the phenomenological point of view. How many of us have moved furniture around to "make more room?"

The WLT showed high overall stability across all levels of analysis. The empirical factor structure matched the hypothesized structure. The reliability and homogeneity of the subscales and overall task were high and the test correlated well with the other Piagetian tasks. In summary the basic results replicate the findings obtained by Pascual-Leone (1970) in a number of unpublished studies.

The same basic results were found for the FIT and ST with the exception of items at the extremes of both tests. The basic results indicated high internal consistency as well as a high degree of relation to the other Piagetian-based tasks.

A major importance of the present research is that it provides support for the possibility of generating developmentally-based scales which are both consistent with Piagetian and psychometric theory. Moreover, the general approach embodied by the "controlled repertoire" procedure would indicate its applicability across diverse populations. A major concern of future research will be to elaborate on the implications of these findings. Furthermore, these results call into question the basic structure of traditional IQ measures. On the basis of these results, it would certainly seem appropriate for future research to take a more detailed look at a large number of traditional IQ instruments, particularly at their use with non-Anglo children.

TABLE 1

FACTOR ANALYSIS - CONSERVATION SCALES
FIVE FACTORS ROTATED
(Principal Component Analysis with Varimax Rotation)

VARIABLE	FACTORS I	II	III	IV	V	h^2
Percent of Variance	22.31	9.67	7.86	6.09	5.06	
Number	.627	.073	.069	-.086	.099	.420
Number	.695	.109	-.009	-.182	.294	.614
Number	.167	.068	.006	-.066	.757	.610
Number	.304	.053	.005	.059	.681	.562
Number	.211	.017	.224	.292	.430	.365
Surface	.154	.131	-.002	.022	.714	.550
Surface	.130	.518	-.082	-.243	.099	.361
Surface	.212	.612	.192	.004	-.050	.458
Surface	-.170	.678	-.008	-.109	.133	.518
Surface	.073	.756	-.062	.106	-.061	.595
Surface	.116	.751	.179	.118	.043	.625
Surface	.028	.733	.208	.166	.167	.636
Substance	.461	.053	.297	.116	.100	.327
Substance	.638	-.032	.206	.114	.109	.475
Substance	.671	.179	.154	.034	.209	.551
Substance	.666	.021	.072	.279	.029	.528
Substance	.576	-.001	.328	.298	.117	.542
Substance	.697	.079	.235	.246	.107	.619
Ego	-.023	-.077	.117	.494	-.091	.273
Ego	.086	.283	.055	.436	.264	.350
Ego	.338	.071	-.329	.571	.055	.557
Ego	.214	-.111	.111	.485	.238	.362
Ego	.149	.050	-.038	.678	-.088	.493
Ego	-.176	.283	.248	.447	.318	.473
Weight	.169	-.085	.664	.148	.314	.597
Weight	.217	-.019	.582	.175	.161	.442
Weight	.097	.302	.540	.070	-.333	.509
Weight	.191	.282	.699	-.033	.016	.606
Weight	.488	.065	.543	.156	.074	.568
Weight	.212	.097	.778	-.162	-.152	.709

TABLE 2

RELIABILITIES FOR THE CONSERVATION SCALES

Method	Number	Surface	Substance	Ego	Weight	N
Cronbach Alpha	.732	.782	.821	.616	.798	106
KR20	.736	.786	.825	.622	.735	106
Homogeneity Ratio	.317	.378	.436	.212	.397	106

TABLE 3

PRINCIPAL COMPONENTS OF WATER LEVEL TASK
THREE FACTORS ROTATED
(Principal Component Analysis with Varimax Rotation)

	FACTORS I	II	III	h^2
VARIABLE		LOADINGS		
Percent of Variance	41.75	13.65	4.49	
Vertical/Horizontal 1	.216	-.839	.[illegible]8	.779
V/H 2	.073	-.725	-.02[illegible]	.531
V/H 3	.239	-.841	.099	.774
V/H 4	.141	-.892	-.023	.817
Tilted 1	.583	-.318	.375	.582
Tilted 2	.609	-.050	.102	.384
Tilted 3	.818	-.098	.037	.680
Tilted 4	.704	-.074	.438	.693
Tilted 5	.810	-.123	.089	.680
Tilted 6	.768	-.197	.223	.678
Tilted 7	.777	-.205	.162	.672
Tilted 8	.785	-.204	-.048	.660
3-D 1	.237	-.061	.594	.413
3-D 2	.358	-.135	.737	.690
3-D 3	-.110	.072	.846	.733
3-D 4	.599	-.331	.387	.618

TABLE 4

RELIABILITIES FOR THE WATER LEVEL TASKS

Method	Vertical/Horizontal	Tilted	Three Dimensional	Total	N
Cronbach Alpha	.820	.902	.696	.627	108
Homogeneity Ratio	.634	.537	.367	.417	108

TABLE 5

RELIABILITIES FOR THE SERIAL TASK

Method	Item One	Two	Three	Four	Five	Six	Seven	Total Scale	N
Cronbach Alpha	-.019	.095	.491	.354	.546	.500	.804	.738	72
KR20	-.000	.095	.502	.368	.557	.512	.812	.729	72
Homogeneity Ratio	-.017	.027	.194	.121	.240	.205	.534	.112	72

TABLE 6

RELIABILITIES FOR THE FIGURE INTERSECTION TASK

Method	Item One	Two	Three	Four	Five	Six	Seven	Eight	Total	N
Cronbach Alpha	.354	.531	.727	.828	.724	.598	.482	.005	.910	91
KR20	.362	.557	.774	.826	.727	.588	.488	.039	.910	90
Homogeneity Ratio	.108	.249	.330	.375	.305	.200	.190	.002	.199	91

TABLE 7

INTER-CORRELATIONS OF MEASURES

ALL SUBJECTS

	Age	Number	Surface	Substance	Ego	Weight	CCS Total	WLT	FIT	ST	D.A.M. MA	D.A.M. IQ	C.M.M. MA
Age													
Number	.327												
Surface	.198	.163											
Substance	.582	.613	.227										
Ego	.423	.317	.313	.397									
Weight	.477	.422	.257	.561	.391								
CCS Total	.568	.693	.526	.817	.663	.767							
WLT	.461	.291	.138	.475	.509	.575	.553						
FIT	.367	.415	.209	.483	.533	.407	.581	.576					
ST	.507	.396	.143	.481	.309	.412	.488	.516	.413				
D.A.M. MA	.522	.332	.040	.330	.471	.306	.426	.486	.330	.430			
D.A.M. IQ	-.392	.003	-.163	-.206	.065	-.134	-.110	.034	-.036	-.044	.565		
C.M.M. MA	.529	.452	.254	.553	.489	.502	.635	.633	.729	.412	.403	-.118	
C.M.M. IQ	-.238	.199	.080	.132	.187	.190	.224	.341	.460	.009	.005	.187	.675

.273 = p .05
.354 = p .01
.443 = p .001

TABLE 8

PRINCIPAL COMPONENTS OF COLUMBIA MATURITY SCALE
LAST 50 ITEMS
FIVE FACTORS ROTATED
(Principal Component Analysis With Varimax Rotation)

	FACTORS					
	I	II	III	IV	V	h^2
VARIABLE			LOADINGS			
Percent of Variance	15.06	5.20	4.82	4.45	4.14	
51	.144	-.053	-.238	-.240	-.121	.152
52	.217	.368	-.116	.097	.001	.205
53	.068	.158	-.054	.574	-.100	.372
54	.227	.154	.072	.525	.047	.358
55	.219	.661	.146	.234	-.177	.592
56	-.054	.399	.659	.016	-.180	.629
57	.075	.118	.336	.121	-.241	.205
58	.104	-.068	.416	.205	-.012	.231
59	-.117	.622	.059	-.036	-.142	.426
60	.212	.134	.207	.403	-.038	.270
61	.017	.096	.378	-.053	-.189	.191
62	.014	.098	.676	-.134	-.014	.485
63	.131	.512	.327	-.105	-.194	.434
64	.271	.327	.143	.068	-.205	.247
65	.300	.138	.019	-.037	-.180	.143
66	-.055	.449	.138	.164	-.040	.252
67	.276	.392	.061	-.120	.084	.255
68	.106	.120	.291	-.286	-.152	.216
69	.226	-.488	.374	-.050	-.303	.524
70	.622	-.024	-.074	.194	-.048	.433
71	.270	.227	-.136	.068	-.398	.306
72	-.018	.284	.028	.030	-.494	.327
73	.422	.071	-.016	-.387	-.279	.412
74	.096	.271	.070	-.577	-.088	.429
75	.570	.167	.209	-.051	-.033	.400
76	.102	.006	.205	.085	-.284	.140
77	.616	-.031	.175	.178	-.194	.480
78	.124	-.250	.314	-.054	-.537	.468
79	.404	.040	.102	.058	-.272	.252
80	.162	.089	.122	-.095	-.346	.178
81	.300	.078	.071	-.367	.029	.237
82	.423	-.022	.278	.364	-.170	.418
83	-.087	.004	.184	-.176	-.558	.383
84	.472	.302	-.112	-.032	-.067	.332
85	.484	-.081	-.099	-.360	-.151	.403
86	.586	.129	.363	-.235	-.056	.550
87	.215	-.076	.420	-.107	.096	.248
88	.517	.093	.313	-.237	-.122	.445
89	.541	.170	.235	.049	-.210	.423
90	.383	.088	-.143	-.232	-.039	.230

TABLE 8 (cont.)

91	.288	-.184	-.171	-.115	-.306	.253
92	.336	-.022	.093	.031	-.100	.133
93	.363	.011	-.132	-.042	-.628	.545
94	.123	.189	.110	-.081	-.351	.193
95	.307	.174	-.180	-.048	-.333	.270
96	.324	-.004	.454	.164	-.160	.364
97	.003	.013	.088	.070	-.566	.334
98	.360	-.040	.108	-.023	-.333	.254
99	.498	-.044	.019	-.014	-.009	.250
100	.214	-.067	.004	.192	-.684	.555

TABLE 9

COMPARISON OF EXPECTED AND EMPIRICAL O-TYPES

	Number	Surface	Substance	Ego	Weight	WLT	ST	FIT
				VARIABLES				
O-TYPE 1								
EXPECTED MEAN	48	63	39	55	45	40	40	40
MEAN	41.471	46.787	38.607	44.578	41.910	42.800	44.390	43.317
STANDARD DEVIATION	11.054	5.742	6.793	4.353	5.261	7.338	6.336	6.770
HOMOGENEITY	.757	.941	.916	.966	.950	.899	.923	.917
t	2.585*	12.310***	0.252	10.442***	2.560*	1.663	3.021**	2.135*
OVERALL HOMOGENEITY OF O-TYPE 1: .911								
O-TYPE 2								
EXPECTED MEAN	48	63	48	55	48	50	50	50
MEAN	54.384	47.719	56.188	48.600	51.878	51.141	50.732	52.201
STANDARD DEVIATION	5.235	8.500	3.847	7.158	8.248	8.259	7.948	8.738
HOMOGENEITY	.951	.865	.974	.906	.872	.870	.875	.858
t	5.977***	8.807***	10.43***	4.38***	4.086***	.676	.451	1.234
OVERALL HOMOGENEITY OF O-TYPE 2: .897								
O-TYPE 3								
EXPECTED MEAN	48	63	48	70	70	60	60	60
MEAN	56.461	62.955	57.789	65.689	61.461	59.418	53.585	59.808
STANDARD DEVIATION	0	10.370	3.154	9.263	5.751	7.545	6.788	8.027
HOMOGENEITY	1.000	.790	.932	.837	.940	.893	.911	.681
t	.00***	.013	9.313***	1.396	4.454**	.231	2.835*	.071
OVERALL HOMOGENEITY OF O-TYPE 3: .907								

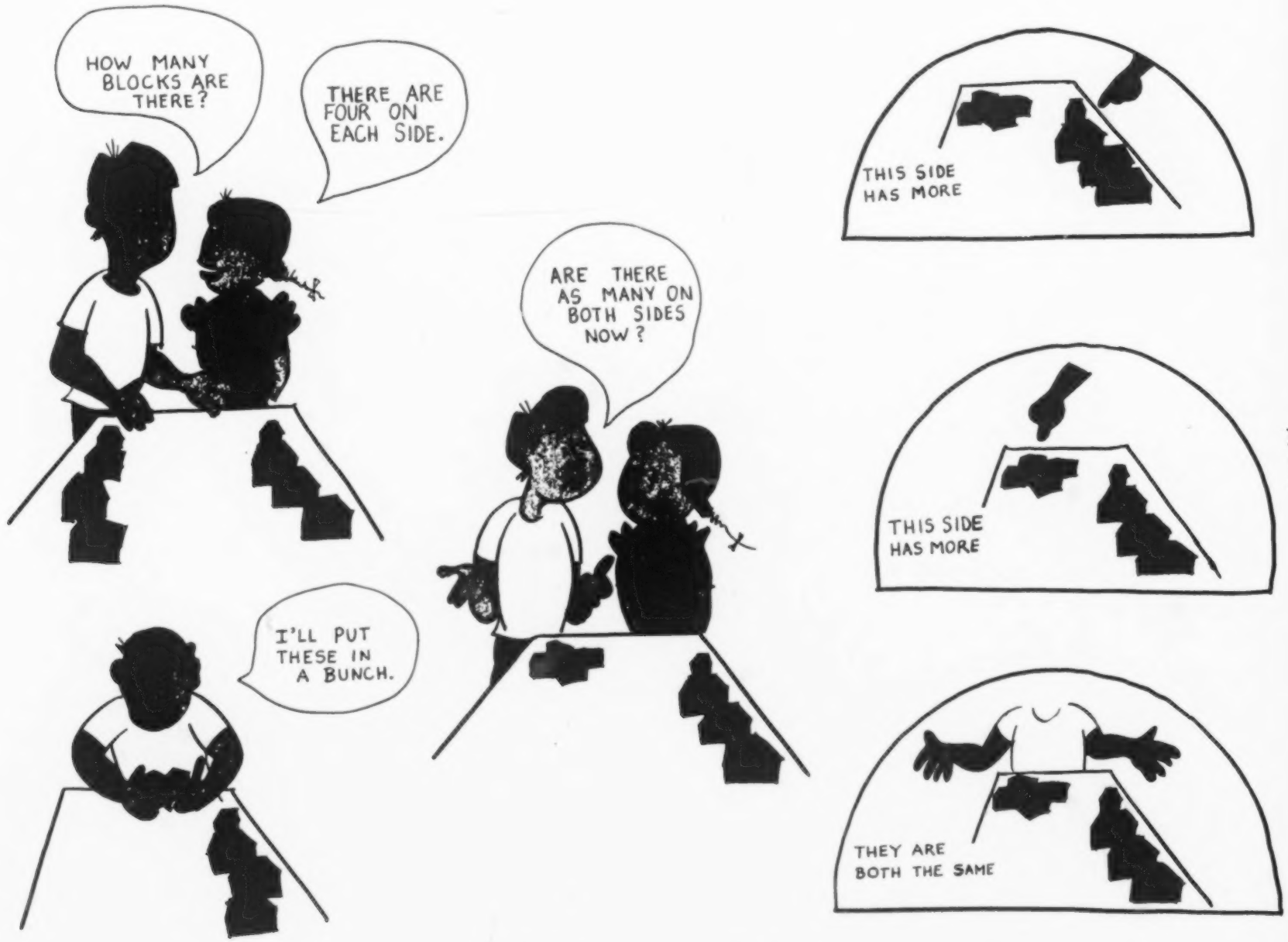
* $p < .05$
** $p < .01$
*** $p < .001$

TABLE 10

COMPARISON OF EMPIRICAL O-TYPES ON SELECTED VARIABLES

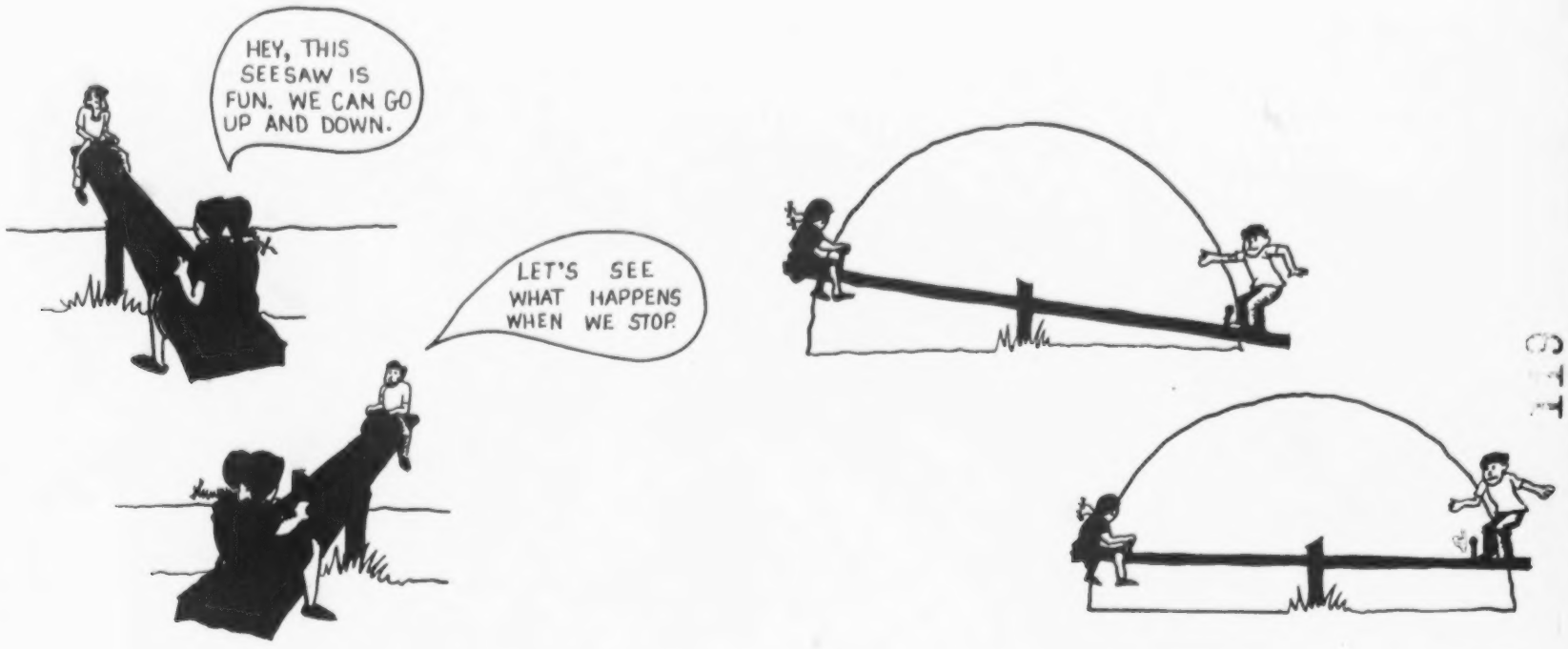
VARIABLE	BDF	WDF	F	O-TYPE 1 vs O-TYPE 2	O-TYPE 1 vs O-TYPE 3	O-TYPE 2 vs O-TYPE 3
Age	2	49	12.071***	3.279**	4.704***	2.294*
Number	2	49	18.254***	5.394***	4.751***	0.681
Surface	2	49	11.813***	1.099	4.715***	3.810***
Substance	2	49	70.587***	10.861***	9.074***	0.813
Ego	2	49	20.150***	1.736	6.077***	5.022***
Weight	2	49	19.105***	3.722***	6.129***	3.452**
Water Level Task	2	49	13.934***	3.377**	5.106***	2.633*
Serial Task	2	49	6.052**	2.787**	3.066**	0.985
F.I.T.	2	49	13.647***	3.537**	4.981***	2.379*
DF				41	26	31

* p .05
** p .01
*** p .001

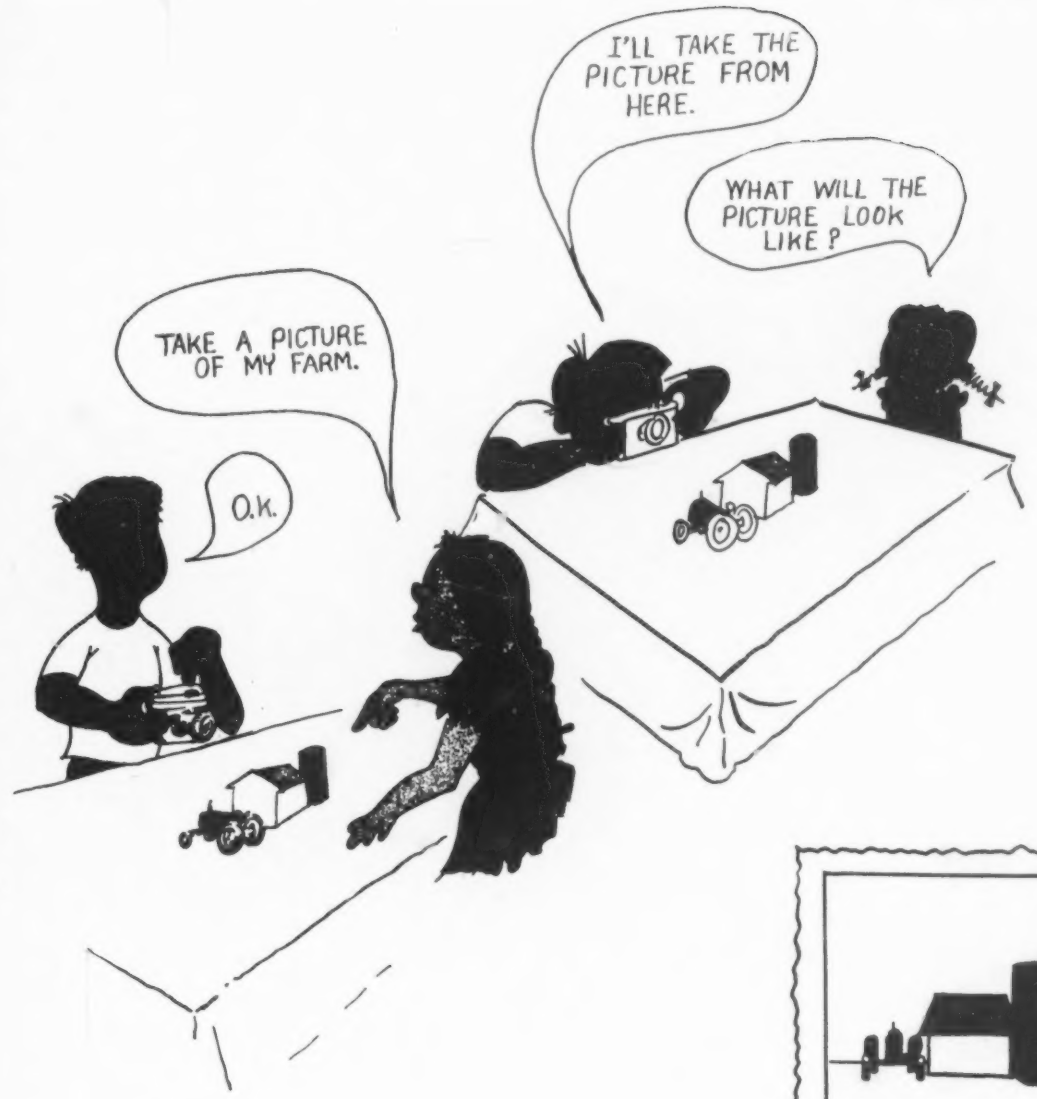

HOW MANY BLOCKS ARE THERE?
THERE ARE FOUR ON EACH SIDE.
ARE THERE AS MANY ON BOTH SIDES NOW?
I'LL PUT THESE IN A BUNCH.
THIS SIDE HAS MORE
THIS SIDE HAS MORE
THEY ARE BOTH THE SAME

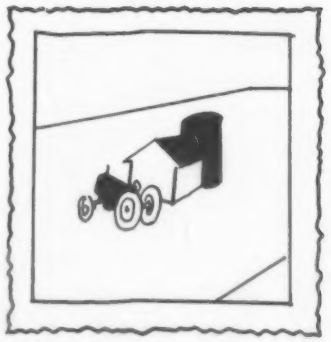
EXAMPLE 1 CONSERVATION OF NUMBER

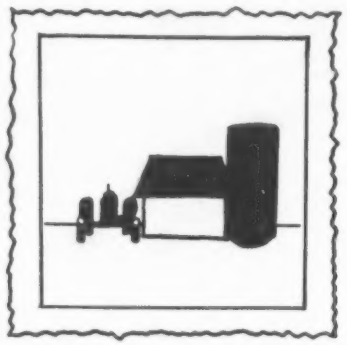
EXAMPLE 2 CONSERVATION OF SUBSTANCE


HEY, THIS SEESAW IS FUN. WE CAN GO UP AND DOWN.
LET'S SEE WHAT HAPPENS WHEN WE STOP.
NOW WE'RE EVEN. WE MUST BE BALANCED.
WHAT WILL HAPPEN IF I STAND UP?

EXAMPLE 3 CONSERVATION OF WEIGHT







EXAMPLE 4 EGO CENTRICITY

REFERENCES

Beilin, H., Kagan, J., and Rabinowitz, R. "Effects of Verbal and Perceptual Training on Water Level Representation." Child Development, 1966, 37, 317-328.

Burgemeister, Bessie B., Blum, Lucille H. Lorge, Irving, "Columbia Mental Maturity Scale", New York: World Book Company, C.1954.

Caldwell, Betty, "Preschool Inventory Revised Edition", 1970 Education Testing Service, Princeton, 1970.

DeAvila, Edward A., "The Use of Artificial Language in the Study of Processing Capacity, Immediate Memory and Piaget's Cognitive Development Variable." Unpublished dissertation, York University, 1971.

DeAvila, Edward A., Randall, David L. and Struthers, Joseph A., "A Group Measure of the Piagetian Concepts of Conservation and Egocentricity", Canad. J. Behav. Sci., 1969 1(4), 263-272

DeAvila E. A., and Phypers, J. M., Boulder Looks at Title I. Denver, Colorado, Colorado State Department of Education, 1968.

Elkind, D., "Children's Discovery of the Conservation of Mass, Weight, and Volume: Piaget Replication Study II." Journal of Genetic Psychology, 1961, 98, 219-227.

Elkind, David, "Conceptions of Intelligence." Harvard Review, 1970.

Harker, W. A. Children's Number Concepts: Ordination and Cordination. Unpublished masters thesis. Dougles library, Quenn's University, Kingston, Ontario, 1960.

Jensen, A. R., "Social Class, Race, and Genetics: Implications for Education." American Educational Research Journal, 1968, 5, 1-42 (a).

Kendler, Tracy S. and Kendler, Howard H., "Experimental Analysis of Inferential Behavior in Children."

Pascual-Leone, Juan, "Cognitive Development and Cognitive Style: A General Psychological Integration." D.C. Heath, New York, in press.

Pascual-Leone, J., and Smith, June. The encloding and decoding of symbols by children: A new experimental paradign and a neo-piagetian approach. Journal of Experimental Child Psychology, 1968, 8, 328-355.

Pascual-Leone, J., DeAvila, E.A., Parkinson, G.M., Goodman, D.R., Development of Cognitive Information Processing. Grant awarded by Canadian National Research Council, 1970.

Pascual-Leone, J., A mathematical model for the transition fule in Piaget's developmental stages. Acta Psychologica 1970, 32, 301-345.

Piaget, J., Inhelder, B., and Szeminska, A., La Geometrie Spontanee de l'Enfant. Paris: P.U.F., 1948.

Piaget, J., and Inhelder, B La Representation de l'Espace Chez l'Enfant. Paris: P.U.F., 1948. English translation: The Child's Conception of Space. London: Routledge & Kegan, Paul, 1956.)

Raven, John C. Raven Progressive Matrices, H.K. Lewis and Company Ltd. London, 1958.

Rebelsky, F. "Adult Perception of The Horizontal." Perceptual and Motor Skills, 1964, 19, 371-374.

Smedslund, J. "The Acquisition of Conservation of Substance and Weight in Children." Scandinavian Journal of Psychology, 1961, 2, 11-20, 71084,

Smedslund, J. "The Effect of Observation on Children's Representation of The Spatial Orientation of a Water Surface." Journal of Genetic Psychology, 1963, 102, 195-201.

Wechsler, David Wechsler Intelligence Scale For Children. The Psychological Corporation, New York 1944.

Witkin, H. A., Dyk, R. B., Faterson, H.F., Goodenough, D. R., and Karp, S. A. Psychological Differentiation, New York: Wiley, 1962.